DOCUMENT RESUME

ED 174 668 TM 009 573

TITLE	Proceedings of the Invitational Conference on Testing Problems. (New York, New York, October 29, 1955).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	29 Oct 55
NOTE	148p.
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Adults; *Aptitude Tests; Educational Counseling; Employment Services; Government Employees; *Occupational Guidance; Occupational Tests; *Predictive Measurement; Psychological Testing; Psychologists; *Testing Problems; *Test Interpretation; Test Results; Vocational Aptitude; Vocational Counseling

ABSTRACT

The conference focused upon the users of tests in counseling and guidance. The first session centered on multi-factor ability test batteries, with papers on Use of Multi-factor Aptitude Tests in School Counseling, by Robert D. North; Use of the General Aptitude Test Battery in the Employment Service, by Pauline K. Anderson; Service Tests of Multiple Aptitudes, by Edward E. Cureton; and Logic of and Assumptions Underlying Differential Testing, by John W. French. Papers in the second session considered methods of improving communication of test information. Particular attention was given to the responsibility of the test user for initiating and maintaining communication with the test author and publisher. Papers were given by John W. Gustad on Helping Students Understand Test Information; Alexander G. Wesman on the Obligations of the Test User; and David H. Dingilian on How Basic Organization Influences Testing. The lucheon address was a re-examination of the role of the psychologist in modern society, presented by Morris S. Viteles. The final session reviewed the relative merits of clinical and actuarial approaches to prediction. Participants in the panel were Nevitt Sanford, Charles C. McArthur, Joseph Zubin, Lloyd G. Humphreys, and Paul E. Meehl. (BH)

****

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

****

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
Educational
Testing Service
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

Proceedings

1955 INVITATIONAL CONFERENCE ON TESTING PROBLEMS

ED174668

TM009 573

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
LOS ANGELES, CALIFORNIA

BOARD OF TRUSTEES, 1955-1956

Lewis W. Jones, *Chairman*

Arthur S. Adams
Samuel T. Arnold
Frank H. Bowles
Charles W. Cole
Donald K. David
John W. Gardner
Henry H. Hill
Frederick L. Hovde
Clark Kerr
Wallace Macgregor
Katharine E. McBride
William G. Saltonstall
George D. Stoddard
Benjamin C. Willis

OFFICERS

Henry Chauncey, *President*
Richard H. Sullivan, *Vice President and Treasurer*
William W. Turnbull, *Vice President*
Henry S. Dyer, *Vice President*
Jack K. Rimalover, *Secretary*
Catherine G. Sharp, *Assistant Secretary*
Robert F. Kolkebeck, *Assistant Treasurer*

COPYRIGHT, 1956, EDUCATIONAL TESTING SERVICE
20 NASSAU STREET, PRINCETON, N. J.
PRINTED IN THE UNITED STATES OF AMERICA
Library of Congress Catalog Number: 47-11220

INVITATIONAL CONFERENCE
ON
TESTING PROBLEMS

OCTOBER 29, 1955

RALPH F. BERDIE, *Chairman*

Multi-Factor Ability Test Batteries in Counseling and Guidance

Communication of Test Information

Clinical vs. Actuarial Prediction

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY LOS ANGELES, CALIFORNIA

FOREWORD

The 1955 Invitational Conference on Testing Problems was concerned with a number of significant and timely topics. Participants, representing diversified backgrounds and opinions in testing, psychology, and related fields, considered the problems involved in the use of multi-factor test batteries in counseling and guidance, methods of improving communication of test information, and the relative merits of clinical and actuarial approaches to prediction. Professor Viteles' luncheon address was stimulating in re-examining the role of the psychologist in our society.

This published record of the proceedings of the 1955 Conference will, I hope, convey to an even greater audience the many new ideas and developments reported and discussed at this conference.

The Chairman of the 1955 Conference, Ralph F. Berdie, met well the challenge of creating an interesting, informative, and successful program. To him and to the speakers I would like to offer our sincere thanks for a job well done.

HENRY CHAUNCEY
President

PREFACE

Users of tests in counseling and guidance make many educational, psychological, and psychometric assumptions, frequently without being completely aware of the nature of these assumptions. The primary purpose of the 1955 Invitational Conference on Testing Problems, sponsored by Educational Testing Service, was to explore and clarify some important principles underlying the counselors' and clinicians' use of tests.

Unlike conferences of immediately preceding years, during which different sessions had been presented simultaneously, this year four consecutive sessions provided an opportunity for the 500 persons attending the Conference to listen to the papers reproduced in this report.

Problems of theoretical and practical importance to users of differential aptitude tests were discussed during the first session. The papers presented descriptions of uses of these tests, descriptions and evaluations of the available tests, and an analysis of some of the assumptions underlying the development and use of differential tests.

Problems of communication among persons concerned with testing were discussed in the second session. Particular attention was given to the responsibility of the test user for initiating and maintaining communication with the test author and publisher. Some refreshing and new points of view were presented from the orientation of an educational administrator.

Participants in the Conference were particularly fortunate to have Dr. Morris Viteles read a paper which analyzed the responsibilities of the psychologist in modern society. The important issues discussed by Dr. Viteles are of interest not only to psychologists, but to all persons whose daily activities bring them into casual or continuing contact with psychologists.

A topic which has received much unsystematic and often heated discussion in the past was reviewed with care by the participants in the afternoon session, the question of the relative efficiency of the clinician and the statistician as predictors. Of particular significance is the paper by Dr. Meehl, which unfortunately because of lack of time, he was unable to read at the Conference.

The success of the Conference was due entirely to the careful work and preparation done by the persons participating on the program and by the very efficient arrangements made by Mr. Jack K. Rimalover and Mrs. Catherine G. Sharp. The chairman also wishes to express his appreciation to the many members of the staff of Educational Testing Service who made this program possible, particularly to Dr. Henry S. Dyer.

RALPH F. BERDIE
Chairman

CONTENTS

FOREWORD by Henry Chauncey 3

PREFACE by Ralph F. Berdie 5

GENERAL MEETING

"Multi-Factor Ability Test Batteries in Counseling and Guidance"

THE USE OF MULTI-FACTOR APTITUDE TESTS IN SCHOOL COUNSELING
Robert D. North, *Educational Records Bureau* 11

THE USE OF THE GENERAL APTITUDE TEST BATTERY IN THE EMPLOYMENT SERVICE
Pauline K. Anderson, *New York State Employment Service* .. 16

SERVICE TESTS OF MULTIPLE APTITUDES
Edward E. Cureton, *University of Tennessee* 22

THE LOGIC OF AND ASSUMPTIONS UNDERLYING DIFFERENTIAL TESTING
John W. French, *Educational Testing Service* 40

GENERAL MEETING

"Communication of Test Information"

HELPING STUDENTS UNDERSTAND TEST INFORMATION
John W. Gustad, *University of Maryland* 51

THE OBLIGATIONS OF THE TEST USER
Alexander G. Wesman, *The Psychological Corporation* 60

HOW BASIC ORGANIZATION INFLUENCES TESTING
David H. Dingilian, *Harbor Junior College, Los Angeles City Schools* 66

LUNCHEON ADDRESS: *"The Psychologist and Society"*
Morris S. Viteles, *Professor of Psychology, University of Pennsylvania* 78

PANEL DISCUSSION

"Clinical vs. Actuarial Prediction"

Nevitt Sanford, *Vassar College* 93

Charles C. McArthur, *Harvard University* 99

Joseph Zubin, *New York State Psychiatric Institute* 107

Lloyd G. Humphreys, *Personnel Research Laboratory, Lackland Air Force Base* 129

Paul E. Meehl, *University of Minnesota* 136

APPENDIX 143

GENERAL MEETING

Multi-Factor Ability Test Batteries in Counseling and Guidance

9

The Use of Multi-Factor Aptitude Tests in School Counseling

ROBERT D. NORTH

Although teachers and administrators in many school systems still rely heavily on general intelligence tests for evaluating academic aptitude, trained counselors are turning in increasing numbers to the use of multi-factor aptitude tests for individual guidance purposes. They look to these tests for the differential measurement of the various aptitudes that are related to academic and vocational success.

Among the multi-factor tests that are currently available for school use are the Chicago Tests of Primary Mental Abilities, the SRA Primary Mental Ability Tests, the Guilford-Zimmerman Aptitude Survey, and the Holzinger-Crowder Uni-Factor Tests. The General Aptitude Battery of the United States Employment Service is offered for school use with high school seniors in some states. This list may be augmented by including multiple aptitude batteries that are based in part upon factor analysis research, or that yield some measures which are essentially in the factor domain, such as the Yale Educational Aptitude Tests, the California Test of Mental Maturity, the Differential Aptitude Tests, and the California Multiple Aptitude Tests. As a group, these batteries provide a coverage of a wide range of factors, extending from those that lie mainly in the area of intelligence to those that represent special aptitudes of vocational nature.

The approach offered by factor analysis—that of measuring human abilities in terms of well-defined primary dimensions—is certainly appealing to the school counselor. However, the practical usefulness of factor scores, as of any other test scores, definitely depends upon their reliability, validity, and normative interpretability. In addition, differential prediction requires that differences among aptitude scores be reliable. Since I presume that more technical discussions of these problems will be presented by some of the other speakers this morning, I shall comment only briefly on these topics.

In regard to the reliability of multi-factor tests, test authors find that they have to strike a delicate balance between maintaining optimal reliability standards and meeting the practitioner's demand for test batteries that can be given within reasonable time limits. For example, the prototype of the current multi-factor tests—the Chicago Tests of Primary Mental Abilities (separate booklet edition)—requires about four hours of administration time for the total battery at the intermediate age level. This administration time is reduced to approximately forty-five minutes, though, in the SRA edition of the Primary Mental Ability Tests at the corresponding age level—evidently to meet the requirements of test users for a shorter battery. While the reliability data given in the Thurstone PMA manuals are not sufficiently adequate to permit a direct comparison to be made of the reliabilities of these two editions, evidence cited by Anastasi (1) indicates that some of the tests in the abridged edition are too low in reliability for satisfactory use in intra-individual measurement.

In order to fit multi-factor tests within practical time limits without sacrificing the needed degree of reliability of measurement, the counselor may find it necessary to restrict measurement to certain selected factors. In the case of most multi-factor batteries, the component tests may be given separately if time limits do not permit the administration of the entire battery. A new measure—the Holzinger-Crowder Uni-Factor Tests—provides for the evaluation of just the verbal, spatial, numerical, and reasoning factors, which are generally found to be more closely related than other factors to academic achievement. Two periods of approximately forty-five minutes each are required for administering this test.

Another approach would be to use a short multi-factor battery for an initial appraisal, and then to supplement this with more intensive measures of certain factors. In other words, it might be desirable to have a multi-factor survey test that would be coordinated with a series of diagnostic factor tests, just as we have survey and diagnostic tests in the reading area. Perhaps multi-factor aptitude test batteries of this type may be published in the future.

It is difficult to make any brief generalizations about the validities of multi-factor tests in connection with their uses in school counseling. We might note, though, that the authors of these tests are tending to give more attention to concurrent and predictive validity, rather than resting their cases entirely on content and construct validity. For instance, a considerable amount of academic validity data is to be found in the comprehensive manual for the Differential Aptitude Tests. It is encouraging, too, to find that some evidence of concurrent and predictive validity for the new Holzinger-Crowder test and the California Multiple Aptitude Tests was gathered in advance of the release of this test for general use.

As the multi-factor aptitude tests become more widely used, the adequacy of the norms will probably be improved. Where only the relative ranking of a student on the various factors is desired, the norms for most of the current multi-factor tests are reasonably satisfactory. But for more precise interpretations of the factor scores, more attention must be given to the representativeness of the norm groups.

For example, one of the important functions served by aptitude tests in the school situation is that of providing an objective basis for evaluating a student's academic achievement in terms of his learning ability. For this type of comparison it is desirable that both the achievement and aptitude test scores be interpretable in terms of the same or very similar norm groups. However, such a condition is not likely to be met in the national norms, except where the multi-factor aptitude tests and achievement tests are prepared by the same publisher, and probably not even then. Establishing local norms is not an entirely satisfactory solution, since some of the advantages of national standardization are thereby forfeited. The large-scale statewide and independent school testing programs have been successful in providing comparable norms on some of the aptitude and achievement tests, but there are many schools that do not fall within the scope of these programs, and hence the norms are not applicable to them. This problem is not unique to multi-factor tests, of course. Let us hope that some solution may be found short of basing norms on Toops' "standard million."

The question of the magnitude of the differences that must be registered among factor scores before such differences may be used as a basis for differential prediction is one with which the school counselor needs considerable assistance from the test technicians. Perhaps the test authors and publishers may develop some improved techniques for helping the test users to understand the necessity for making allowances for the standard error of score differences in the interpretation of test profiles. A noteworthy forward step in this direction has been taken by ETS in connection with the profile charts that have been prepared for the new Cooperative School and College Ability Tests. Discussions in the test manuals of the principles of standard errors are often helpful to the school counselor if the terminology that is used is not excessively technical.

Turning now to more general considerations concerning the use of multi-factor tests in school counseling, one of the questions that might be raised is whether these tests may be employed effectively at the elementary school grade levels. On the whole, there seems to be very little evidence that factor scores have any practical advantages over general intelligence test scores at these lower grade levels. In the negative direction, Garrett's (2) findings indicate that intelligence factors are relatively undifferentiated among young children, and that mental abilities do not tend to become specialized until adolescence or early adulthood. However, as Vernon (3,pp. 29-31) points out, the relation between the pattern of mental organization and chronological age is not clear-cut, and the research data in this area are often difficult to interpret because the variables of group heterogeneity and appropriateness of test content for different grade levels are not adequately controlled. At the present stage of development of multi-factor tests, though, it seems advisable that considerable caution be observed in using these tests at the elementary school level, and that careful attention be given to the correlations among the factors and the degree of reliability of the differences among the scores.

Regardless of the grade level at which the multi-factor tests are used, it is essential that the counselor keep in mind the difference between the selection and guidance applications of the results. When tests are used for selection purposes, the principal objective is to evaluate the individual's present aptitudes in order to predict his probable academic or occupational success. In this case, relatively little consideration is given to the possibility of improving the individual's aptitudes when they are low. In the guidance use of the test results, on the other hand, it is not only the individual's present aptitudes that are important, but his potential for development as well. If a low aptitude score may be a reflection of lack of opportunity for development, some attention should be given to the possibility of encouraging the improvement of this aptitude through guidance and instruction.

For example, suppose an individual has a high verbal aptitude score and a low numerical aptitude score. Does this mean that the student should be guided toward school courses that emphasize verbal skills, and away from courses involving mathematics? Or should the student be stimulated to improve his numerical ability? A satisfactory answer to this question probably requires more information about the underlying causes of the differentiation of mental abilities than is presently available. In the absence of any conclusive evidence to the contrary, it would be well to keep in mind Vernon's point of view that "factors over and above g arise, partly perhaps from hereditary influences, but mainly because an individual's upbringing and education imposes a certain grouping on his bonds" (3, pp. 31-32). In any event, it is important that counselors recognize that they must make certain assumptions about the determinants of intra-individual differences in mental abilities when they counsel from profiles.

With the attention that is now being given to multi-factor aptitude test scores, one might ask whether there is any value in the general intelligence test score, or intelligence quotient, in the school situation. It may be recalled that when the Chicago Tests of Primary Mental

Abilities were introduced, the Thurstones stressed that general intelligence scores are of little value and that only factor scores should be used. They did not provide for an over-all intelligence quotient on the original edition of the test. However, they soon found that when school teachers and counselors use an intelligence test, they expect to get an intelligence quotient from it. The SRA Primary Mental Abilities Tests now provide, in addition to factor scores, a total score that is designed to serve as a single index or average of the child's intelligence level.

Whether the IQ is used as a measure of **g**, or merely as a summary of the pupil's performance on the test as a whole, it does seem to have some practical utility. It provides a basis for sectioning pupils, where this must be done on an over-all academic ability basis; it serves as a general guide for estimating the desirability of encouraging a pupil to prepare for a college career; and it is a key in vocational counseling to the general occupational level for which the student should aim.

Of course, if a general intelligence score is desired for purposes such as these, it may be obtained from either an omnibus test or from one of the multi-factor tests that yields a composite score in addition to factor scores. In the latter case, it might be desirable for counselors to know the effective beta weights of the factors, instead of just the raw score weights that are used for computational purposes, so that they might be in a better position to interpret such discrepancies as may be found among the total scores of a single individual on several multi-factor tests.

In summary, the multi-factor tests apparently are beginning to meet the needs of the school counselor for a means of making differential predictions of academic and vocational success. The practical usefulness of these tests in the school situation will increase as the reliability, validity, norms, and theoretical framework of the factor scores become more adequately established. Meanwhile, the general intelligence score continues to have an important role in school counseling, particularly at the elementary school grade levels.

REFERENCES

1. Anastasi, Anne. An empirical comparison of certain techniques for estimating the reliability of speeded tests. *Educational and Psychological Measurement*, 1954, 14, 529-540.

2. Garrett, H. E. A developmental theory of intelligence. *American Psychologist*, 1, 372-378.

3. Vernon, P. E. *The Structure of Human Abilities*. London: Methuen & Co., Ltd., New York: John Wiley & Sons, Inc., 1950.

The Use of the General Aptitude Test Battery in the Employment Service

PAULINE K. ANDERSON

A national system of public employment service offices administered by the states, but financed, coordinated and given general technical supervision by the Federal Government was established in 1933 by the Wagner-Peyser Act and has been in continuous existence since that time. The Agency of the Federal Government which supervises and coordinates the State Services is the United States Employment Service of the Bureau of Employment Security of the U. S. Department of Labor.

Among the functions for which State Employment Services nationally are responsible are, of course, placement itself; i.e., providing the right worker for the right job at the right time and secondly for the vocational counseling of those who need help in choosing suitable fields of work or help in resolving a wide variety of job adjustment problems. The testing program of the Employment Service which consists of both aptitude and proficiency tests is the result of continuous research since the establishment of the agency—research which has involved the cooperation and participation of workers, employers, unions, schools and colleges throughout the country. The General Aptitude Test Battery which, as I am sure you know, is a multi-factor test, is used primarily in connection with our counseling program, especially with young workers just entering the labor market. But it is used successfully also in the counseling and placement of other applicant groups, particularly veterans, older workers and displaced workers. Like any other test used by the Employment Service, the GATB is a technique by which we attempt to make our counseling and our placement more accurate and more effective. It is an integral part, in other words, of our total service to both applicants and employers; it is used only where it is needed; its results are interpreted in the light of total pertinent information about the individual and about jobs.

As an Employment Service our primary interest is of course in the occupational qualifications of our applicants. Our tests therefore are designed to measure qualifications for occupations as these have been determined by experimental evidence secured primarily, though not exclusively, from samples of workers performing successfully in the particular occupations for which we have developed test norms. Our aptitude test battery, for example. for Boarding Machine Operator, a job found in hosiery manufacturing, consists of a combination of single tests which in combination have been found to be effective in discriminating between better and poorer employed Boarding Machine Operators. This is an aptitude test battery which our offices might utilize to select an applicant inexperienced in the occupation for referral to an employer who had placed an opening with us for a Boarding Machine Operator trainee.

However, those of you who are counselors know that when you are dealing with a person who needs help in choosing or confirming a vocational goal you cannot start with the requirements of a specific job opening. You must start rather with the individual himself and try to appraise his vocational aptitudes as broadly as possible and relate these aptitudes to the requirements of jobs. As many of you know also there are quite literally thousands of specific jobs, in fact some 25,000 have been defined by the *Dictionary of Occupational Titles*. Fortunately however, many of these can be grouped into job families on the basis of various kinds of similarities. Such groupings make it possible and practical to help an applicant select broad areas or fields of work in which he has chances for successful performance, within which there are a large number of specific jobs any one of which may serve as a starting point for him. The GATB, because of its nature, allows the counselor to do exactly these two things; that is, explore applicants' abilities broadly and relate them to the aptitude requirements of fields of work established on the basis of the similarity of their aptitude requirements. The battery, consisting of 12 single tests 8 of which are paper and pencil and 4 of which are apparatus, measures nine vocational aptitudes which have been found to be of most significance in most jobs occurring in this country today and relates the individual's aptitude scores in each of these nine to the aptitude requirements of many broad fields of work which include well over 3,000 specific occupations. This amount of occupational coverage is secured from a group test session which lasts approximately 2½ hours. The aptitudes the battery measures are general intelligence or scholastic aptitude (G), numerical (N), verbal (V), and spatial (S) aptitudes, clerical perception (Q), form perception (P), motor-coordination (K), and finger and manual dexterities (F and M). These aptitudes originally were identified by means of factor analysis studies which involved the administration of 59 single tests to a sample of over 2,000 individuals. The fields of work for which the battery scores represent a wide range of type and skill level ranging from semi-skilled machine tending and machine operating, simple inspection and routine clerical work through skilled machining, printing and bookkeeping, to professional nursing, teaching, accounting, engineering, etc.

The battery's results are expressed in two forms: The Individual Aptitude Profile; that is, the aptitude scores obtained in each of the nine aptitudes measured. This profile gives a numerical representation of the applicant's own strengths and weaknesses as well as his strengths and weaknesses as compared with the general working population of this country. General population norms were established on the basis of a stratified quota sample of 4,000 workers selected from over 8,000 cases reflecting an exact representation of the occupational distribution of the national labor force as given by the 1940 census except for certain deliberate exclusions of farm, forestry, mining and personal service occupations. In addition to its occupational representativeness the sample is typical of the age, sex and geographical distribution of this country's working population. The mean for each aptitude score has been set at 100 with a standard deviation of 20. Thus the counselor is able to see by a glance at the Individual Aptitude Profile whether the applicant generally tends to meet, exceed, or fall below this average. He can see at a glance also in what kinds of aptitudes he tends to excel or fall below; that is, the cognitive vs. the motor, the more abstract vs. the simpler and more concrete, etc. The profile also makes it possible to see quickly where the applicant's own best abilities lie regardless of his standing in relation to the general population. Thus it can be seen, for example, that John Jones does best in numerical and spatial aptitudes and poorest in verbal aptitude and clerical perception.

As I indicated earlier, however, our main concern is with the occupational qualifications of each applicant. Therefore in day to day work with the GATB, the counselor's main interest is centered less in the Individual Aptitude Profile than in the fields of work for which the applicant qualifies. The GATB indicates such fields by means of Occupational Aptitude Patterns which consist of jobs grouped together on the basis of the similarity of their aptitude requirements. Those jobs which have been found to require the same combination of significant aptitudes to the same minimum degree constitute an Occupational Aptitude Pattern. Each pattern uses the multiple cut-off method to determine occupational qualification or non-qualification at least on the basis of test performance. Thus, the individual is considered qualified for an Occupational Aptitude Pattern only if he meets the minimum score on each of the aptitudes found to be significant for this particular family. The aptitude requirements themselves have been established on the basis of experimental evidence secured from samples of workers employed in the occupations making up the field or Occupational

Aptitude Pattern plus, in some instances, samples of senior students and/or apprentices successfully completing particular courses of training for certain occupations on the vocational, technical or professional level.

One of the most significant contributions made by the GATB lies in the fact that it helps to underline what most of us know but too frequently tend to forget in practice; namely, that most people can do more than one thing well and that a high degree of what we call general intelligence or scholastic aptitude is of primary importance in only a rather small percentage of the total number of existing jobs. Thus, the GATB brings out the fact that even for persons of rather limited intellectual ability there exist many occupational outlets in which their performance can be *not marginal* but truly successful. For example, here is a high school graduate whose G, V, N scores are 93, 90, 90 but who still qualifies for seven fields of work two of which consist of many kinds of clerical jobs. Here is another case of a girl who probably is defective—her scores are 67, 68, 67, and 78 in G, V, N, and S respectively but who still qualifies for four different fields of work, one of which in itself includes literally hundreds of semi-skilled industrial occupations involving machine tending and operating.

The GATB also of course helps uncover those who could profit from higher education. We encourage such people to consider college when other evidence also supports the evidence of the test, and frequently through referral to and assistance from other community agencies make it possible for them to go to college. One of our offices recently tested a high school graduate who had no thought of going on to college. He expressed interest in clerical jobs but his G, V, N, and S scores were 143, 129, 141 and 130 respectively. He qualified, among many other things for both accounting and engineering. College as a possibility was discussed with him and he was referred to our Consultation Service for specific information about individual colleges and their requirements and for possible financial assistance. In the meantime he was placed on a summer job as a file clerk.

Another applicant was a 24 year old Korean Veteran, a high school graduate whose only civilian work experience had been as an order filler—an unskilled job. His assignment in the Army was as a general clerk. Some of his GATB scores were:

G	V	N	S	
137	131	122	130	etc.

He qualified for many professional and technical occupations. His Interest Check List also indicated many scientific preferences. He too was referred to our Consultation Service for assistance in college planning and he too was placed—as a Chemist Assistant.

Thus the GATB by its nature makes it possible for us to help the applicant choose vocational goals which will utilize his maximum potentialities. And this is a very fundamental policy of both our counseling and placement activities—to make maximum utilization of the potential or actual skills of our applicants. An applicant, for example, may qualify according to the test results for a field of work which includes occupations on a highly skilled trade level; e.g., machinist, tool and die maker, etc. He may qualify also for a variety of machine operating and tending jobs which are mostly of a semi-skilled nature. Unless there was some very good reason having to do with the individual himself, the counselor of course would encourage the more skilled occupations as the goal. Or an applicant may qualify for an Occupational Aptitude Pattern which utilizes only one or two of his own best abilities and qualify also for a second which utilizes more of his own aptitude strengths or utilizes them at a higher level. Again unless there was other specific evidence against it the counselor would encourage consideration of the second field, even if the skill classifications of the jobs in each were equal. Or, to take a converse situation, an applicant may express some interest in engineering as a vocational goal but we may find that, at least as far as the test results are concerned, he meets the minimum requirements for machinist but does not come anywhere near meeting the minimum requirements for either drafting or engineering. In other words the GATB helps indicate the uppermost level of skill within related fields of work to which the applicant may aspire.

Obviously, when we use the term *maximum* utilization we use it in a relative not an absolute sense. Sometimes an applicant may just barely meet the minimum requirements of a field but show greater strength for a related field on a slightly lesser level of skill. It may be that for this individual maximum utilization would be achieved more successfully through the somewhat lower level jobs. This brings me to the most important point that I wish to leave with you, one which I mentioned earlier; namely, that we use test information as only one piece of information, important though it may be, about a total individual who is not after all made up just of aptitudes nor whose job success will be based solely on aptitudes. His interests, his goals, his education, his work history if any, all must be taken into account in order to determine accurately his total occupational qualifications and to make relatively accurate predictions of his probable job performance. In the Employment Service we use the short hand *SKAPATI* to emphasize the variety of factors which must be evaluated in order to arrive at a sound occupational classification which in turn will lead to accurate matching of the applicant's qualifications with job requirements.

SKAPATI is translated as follows: The "*SKA*" equals *s*kills, *k*nowledge and *a*bilities—information about which is secured primarily through interviewing the applicant about his school and work record supplemented, where necessary, by an actual school record or by a check with former employers. The "P" of SKAPATI stands for physical capacities, that is, the general physical condition of the applicant and/or any physical disabilities which he may have that need to be taken into account in helping him find a suitable field of work. Information about physical capacities is secured through interview information, observation and where necessary by medical reports. The second "A" is for aptitude information which is secured through interview information supplemented by aptitude test measurement. "T" and "I" stand for personal traits and interests, information about which are secured again primarily through the interview and observation plus the use of an Interest Check List and/or reports from schools, former employers, social agencies, etc. Since all this information is given weight in the counseling situation there are instances in which an applicant might very well be encouraged to work toward a goal even though his test scores, when looked at in isolation, might indicate less chances of successful performances. Of course in such cases the counselor would indicate to the applicant that he might have to work a little harder than others to achieve success or that he might have to be satisfied with satisfactory rather than outstanding performance. But the important point is that aptitude qualifications by themselves, even though derived from an instrument as well standardized as the GATB, nevertheless are never used as the sole basis for counseling or for placement. In fact our total counseling process and our use of tests in it are rooted in an individual, analytical approach whereby the counselor attempts to recognize and understand the particular individual's uniqueness as an individual and to help him realize this through both his choice of work and his actual entry into the field of his choice.

Service*Tests of Multiple Aptitudes

EDWARD E. CURETON

The present trend in aptitude testing, both for educational and vocational guidance and for employment and placement, is away from both general intelligence testing on the one hand, and from measurement of specific vocational aptitudes on the other. A profile of aptitude scores provides more information than does an average, and even in the cognitive area there are important measurable aptitudes which do not fit the usual definitions of general intelligence. But the development of a special battery for every important occupation is a hopeless task, and even if we had such a set of batteries, no examinee could possibly take all of them. The results of the factor analysis studies made during the last 20-odd years point the way to a reasonable compromise. It appears from them that a battery of perhaps two or three dozen tests will measure most of the important cognitive, perceptual, and sensori-motor aptitudes, and that a battery of only a half-dozen to a dozen will measure the really crucial ones fairly well. For at least a large number, if not a majority, of the thousands of different occupations, the best prediction of success yielded by such a battery will not be improved greatly by the addition of special tests, so long as these latter still measure cognitive, perceptual, and sensori-motor aptitudes. Tests of interest, attitude, personality, and the like are another matter, but in these domains much work remains to be done in test development before the procedures of factor analysis can be expected to yield anything approaching definitive results.

This is not to say that in the domains here at issue—the cognitive, perceptual, and sensori-motor—the factor analysis results are as yet definitive, but in these areas they are at least highly suggestive, and there is enough agreement among them to permit some useful application. The trend toward such application has already appeared, and the purpose of this paper is to assess its present status.

My first and most important task was to try to identify and characterize in some uniform language the factors measured by the tests of sixteen batteries. The material at the beginning of Table 1 lists the

---

*By a service test, we mean a test designed for service uses such as guidance, employment, placement, etc., as contrasted with the experimental batteries used in most factor analysis studies. Batteries developed by the military services are *not* included in this review.

factors and abilities employed. The characterizations themselves are in the body of the table under the column heading, "Probable Factors." This was strictly an arm-chair job. The lists of factors and abilities are certainly incomplete. By any reasonable definition of a factor, some of those listed are too broad, and one or two may be too narrow. Many of the characterizations are undoubtedly in error, even in terms of these lists.

Whenever it appears that a test should have high loadings on several important factors I have listed several, in the order of their presumed importance or magnitude from left to right. There will be more errors, needless to say, in these judgments of order than in the judged factors themselves. Despite all their obvious shortcomings, however, I venture to hope that these characterizations will be sufficiently valid for most practical purposes until such time as we can conduct a large factorial study of several of the service batteries themselves. Such a study would also yield equivalent scores for the various batteries, and this also is so important that when such a study is made I am not sure which objective would be primary and which secondary.

In deriving the list of factors I was fortunate in having available the work of a set of committees which, under the auspices of the Educational Testing Service, identified sixteen aptitude factors as fairly well established, and named at least three tests which could be recommended as reference tests for each in future factor analyses. For their objective, the preparation of lists of standard reference tests, two factors could be considered distinct whenever they could be defined in terms of two distinct groups of reference tests. In selecting reference tests, moreover, they could occasionally include one with a known high loading on the given factor but a higher loading on some other factor. But for my objective, the factorial characterization of tests, some of which have not appeared in actual factor analyses, a more rigid criterion for a distinct factor seemed essential. First, it must be distinct from all others in terms of explicit concepts, not merely in terms of different sets of reference tests. Second, in terms of these concepts it must be broad enough to imply at least two or three reference tests which differ in apparent content. On applying these criteria to the committees' sixteen factors, it appeared necessary to modify some of them for my purpose.

Though I agree with the committee that vocabulary is the central core of the verbal factor, I broadened it to include such tests as paragraph meaning, proverb matching, sentence completion, verbal analogies, and even general information. The general concept used might be termed verbal comprehension. There appears to be good evidence in the factor analysis literature for the existence of a verbal factor, or rather a group of verbal factors, covering this range. I am not sure, on the other hand, that the tools of English expression, such as spelling, grammar, punctuation, and the like, belong under this factor, and in fact their factorial structure seems not well determined so far. I listed them, therefore, as two abilities: spelling separately because it is usually tested separately, and the others under the general term, "language usage," because they are frequently measured together. By an ability I mean merely an area of measurement which is conceptually distinct but whose factorial structure is as yet not well determined.

The deduction factor was retained, but the concept was narrowed to include only formal tests of the syllogistic type. This comes perilously close to violating my own criterion of breadth, but the factor analysis results required its retention. Two of the committee's reference tests were of the syllogistic type, but the third was a verbal analogies test, which, in those factor analyses with which I am acquainted, loads substantially on the deduction factor but higher on the verbal factor.

From such knowledge as I had of factorial literature, including the committee's report, I was unable to formulate clearly distinct concepts for the reasoning factors other than deduction, so I lumped them all together under one heading. The same situation appeared in the case of the space factors and the fluency factors, so each of these groups was represented also by a single factor.

To cover certain other tests in the sixteen batteries, I added an immediate memory factor to the committee's list. This also undoubtedly represents several factors, but we do not yet know how to differentiate among them.

The committee listed only one perceptual speed factor, and this I let stand despite the evidence from the USES studies that clerical speed may be a separate factor.

To characterize one or two tests in the sixteen batteries, I postulated a factor, visual discrimination, which so far as I know has not been found in any factorial study. I am quite certain it *will* be found as soon as appropriate tests are included in such studies. Appropriate tests would include sets of lines, not parallel, with one just slightly longer or shorter than the rest to be identified; sets of arcs of circles with one having a just slightly longer or shorter radius than the others, or one which is not quite circular; sets of circles with smaller concentric circles within them and one not quite concentric; and the like. These tests, unlike the identical forms tests which measure the perceptual speed factor, would be administered with generous time limits. They would, I predict, reveal a factor representing individual differences in a more or less general visual difference limen. Drake has already shown, for the concentric circles test, that it is a valid predictor of effectiveness in several types of visual inspection work.

Flanagan has argued with some cogency that reasoning and judgment are not the same, and the concepts as such are discriminable. So I added judgment as another unanalyzed ability, of which there are probably several distinct varieties.

It was necessary to add two other abilities, mathematical achievement and science information, to cover tests in one or two service batteries. There have been some factor analyses of achievement tests, but the results of these studies do not yet seem to me to warrant confident identification of the factors generated by the study of school subjects above the elementary level.

Finally, the characterizations were limited to factors and abilities measured by paper-and-pencil tests. Two of the service batteries include apparatus tests, but these tests and the factors generated by them do not appear in Table 1.

Wherever service batteries were constructed directly from factor analysis data, the characterizations of them in Table 1 will usually agree substantially with those of their authors. Some exceptions will occur because of my use of combined factor-categories, and a few more due to differences in sheer nomenclature. Thus, my concept of the number factor is essentially speed and accuracy in solving problems of *no* intrinsic difficulty, but some authors use this term to cover the harder tests of arithmetic computation and problem-solving. In Table 1, such tests are characterized as reasoning tests.

The sixteen service batteries fall roughly into three categories. In the first we find those whose primary objective in test selection seems to be to approximate pure-factor measurement as closely as possible. In this category we find the first nine tests in Table 1:

ACE Primary Mental Abilities,
Chicago Primary Mental Abilities, ages 11-17,
SRA Primary Mental Abilities, ages 11-17,
SRA Primary Mental Abilities, ages 7-11,
SRA Primary Mental Abilities, ages 5-7,
Holzinger-Crowder Uni-Factor Tests,
Factored Aptitude Series,
Guilford-Zimmerman Aptitude Survey, and
USES General Aptitude Test Battery.

In the second category we find batteries developed on a compromise basis. Range of factorial coverage was an objective but factorial purity of the tests was less important, and the particular tests were designed to be similar to others which had been shown to be valid predictors of a variety of educational and occupational criteria, or to measure abilities shown by job analyses to be important elements of many occupations. This category includes batteries IX through XII in Table 1, namely:

USES General Aptitude Test Battery,
Differential Aptitude Tests,
Multiple Aptitude Tests, and
Flanagan Aptitude Classification Tests.

The USES General Aptitude Test Battery is a transitional type listed in both the first and second groups, and the Flanagan Aptitude Classification Tests are listed in both the second and third groups. In the third category we find batteries designed to predict multiple criteria, often over only a limited range such as educational curricula, but sometimes over a very wide range. An attempt, necessarily somewhat hasty and hence probably less than completely successful, was made to include here all batteries belonging properly in the first two categories, but excepting those whose distribution is restricted to the armed forces, other government agencies, private firms, and organized testing programs. In the case of the third category, no such attempt was made. Table 1 includes only the tests in this category numbered XII through XVI:

Flanagan Aptitude Classification Tests,
Yale Educational Aptitude Test Battery,
Aptitude Tests for Occupations,
Engineering and Physical Science Aptitude Test, and
Cleeton Vocational Aptitude Examination.

These are merely a small group of such batteries—the ones which came to hand most readily during the preparation of this paper. They were included merely to show that the multiple validity approach and the multiple factor approach to battery construction often lead to quite similar productions. Factor analysis provides a clearer rationale, however, and on the practical side it shows which valid tests can safely be omitted from a battery if certain others are present in it.

To report in any detail the data on the reliabilities, validities, and norms of these batteries is impossible within the limitations of a short paper, and valid comparisons are impossible in the absence of data from several batteries given to the same group. And since computing methods as well as reporting methods vary from manual to manual, there seems to be no useful method of summarizing such information in tabular form. My remarks on these matters must therefore be brief, impressionistic, and somewhat scattered.

For Thurstone's original ACE Test for Primary Mental Abilities, I had only copies of the booklets, but it is my impression that this battery is not very widely used for service testing, and that most of the studies in which it has been used were experimental in nature and are reported in journal literature. Its immediate successor is the Chicago Tests for Primary Mental Abilities, but this battery is an easier version based on a second factor analysis, and is recommended only for children ages 11 to 17. The data here and in Table 1 are for the single-booklet edition, a shorter version of a previous multiple-booklet, separate-answer-sheet edition. The single-booklet edition is booklet-marked. Percentile norms for the six factor scores are provided for each half-year age group from 11 to 17.5, based on the scores of about 18,000 Chicago children in 29 elementary and 31 high schools. A profile chart is printed on the back cover of the test booklet. High correlations between the factor-score composites and the actual primary-factor scores are reported, along with the results of a second-order analysis of the correlations between the primary factors. This latter analysis yielded a single general factor, with the highest loading on reasoning, fairly high loadings on verbal, fluency, and number, and lower loadings on space and memory. Reliabilities are given for the test scores and the factor scores, but no correlations with external criteria are reported.

The SRA Primary Mental Abilities for ages 11 to 17 is a shorter version of the same battery, with the memory tests omitted and the remaining five factors measured by one test each. The test booklet is re-usable, with a replaceable carbon-back answer sheet, and there is also a machine-scored edition. Percentile norms for each year of age from 11 to 17-or-over are incorporated directly into the profile chart. They are based on the scores of those children of each given age who were in junior and senior high schools, *not* on the full age-ranges. The original norms were apparently developed from the data for the parent battery, but further adjustments were made on the basis of a second sample whose size was not stated. Deviation-type quotient norms are provided also on the profile sheet, with mean 100 corresponding to percentile 50, and standard deviation 16. The reliabilities and intercorrelations of the tests are reported, as are also their correlations with several intelligence tests, the Iowa Tests of Educational Development, the Stanford Achievement Test, and the USES General Aptitude Test Battery. For these last data, for a tenth grade sample and a twelfth grade sample, with the Minnesota Clerical Test and the Revised Minnesota Paper Form Board Test included also, two factor analyses are reported. The factors found were intelligence (actually reasoning), paper-motor speed (a fusion of our aiming and motor speed factors), space, perceptual speed, number, dexterity (from the four apparatus tests scores), and verbal. No correlations with external criteria are reported.

The SRA Primary Mental Abilities for ages 7 to 11 is the only multiple aptitude test available for use with this age range. The test booklet is re-usable, with a carbon-backed answer sheet. The accompanying materials provide age norms for ages 6-0 to 14-0 at two-month intervals for the five factor scores and for the four separate tests of the verbal and reasoning factors, based on scores of 4,744 children aged 7 to 12, and revised on the basis of 2,000-odd additional cases. The norms for ages 6 to 7 and 12 to 14 are extrapolated. IQ and non-reading IQ estimates are obtained from weighted composites of five and three test quotients respectively, and directions are provided for computing and using a reading aptitude quotient, an arithmetic aptitude quotient, and a measure of current reading/experience status based on the differences between scores on the written and picture-oral tests of vocabulary and reasoning. An interpretation booklet includes a profile chart embodying the age norms, and directions for making the IQ, reading, and arithmetic estimates. A technical supplement to the manual reports the reliabilities and intercorrelations of the test scores, and correlations with several intelligence tests, reading tests, and arithmetic tests.

The SRA Primary Mental Abilities Test for ages 5 to 7 is the only multiple aptitude battery available for use with young children. It is entirely oral and pictorial, and is booklet-marked. The cover page of the booklet includes a profile chart which embodies age norms from 3-0 to 9-0 at two-month intervals, and a separate column of age scores for a weighted composite total score representing intelligence, based on scores of 1,200 children aged 5 to 8. The manual discusses methods for estimating reading readiness, arithmetic readiness, and motor coordination. The technical supplement reports reliabilities and intercorrelations among the scores, and correlations with the Stanford-Binet, Form L, and with several reading, arithmetic, and general school achievement tests. In some cases the achievement tests were given more than a year later. The evidence presented justifies the claim that this test provides fairly valid evidence of first-grade readiness.

The Holzinger-Crowder Uni-Factor Tests appear in two equivalent machine-scorable forms. End-of-year percentile norms for each factor score are given for grades 7 to 12, and also for the scholastic aptitude score. A profile chart is provided on the back of one of the three answer sheets. Age norms are not given in the manual, but a footnote says they can be obtained from the publisher. The norms are based on the scores of over 10,000 students in grades 6 through 12 in 38 schools in 28 communities in 7 states. An IQ table for the Scholastic Aptitude score is provided also, based on equi-percentile equating to the Terman-McNemar test in a subsample of over 2,000. Alternate-form as well as split-half reliabilities are reported, along with the intercorrelations of the factor scores. Correlations are reported between the factor scores and several dozen achievement tests, a considerable variety of high school subject grades, and seven intelligence tests, but none with external criteria of occupational success. Data are also given on sex differences: boys are better on space, girls on verbal and scholastic aptitude, and the results for number and reasoning are inconclusive, though the girls have again a slight edge. Practice effect is substantial for space, slight for reasoning, and negligible for verbal and number. The essential interchangeability of the two forms is shown to justify one set of norms for both.

The Factored Aptitude Series consists of sixteen four-page test booklets, and a seventeenth to be used for recording the results of a nuts-and-washers apparatus test. The tests are all booklet-marked. On the back of each test booklet is a brief rating scale: "On a regular job I would like to do the kinds of tasks represented by this test: A) As a major part of my work, B) Frequently, C) To a moderate degree, D) Only occasionally, E) As seldom as possible. Remarks................" An interest index for job areas is based on the ratings for the several tests, and correlations with the Kuder, Lee-Thorpe, and Strong are reported to range from .35 to .70. Reliabilities, intercorrelations among the tests, and validities against job criteria are reported in the manual and in various published notes, usually without exact data on the sizes and compositions of the samples. A complete statistical report is in preparation, but was not available at the time this report was written. Stanine and percentile norms are given for the seventeen tests, based on samples of the working population of unstated size, but apparently quite large. Tables are given for converting stanines on sub-batteries into weighted aptitude indices for 24 basic job-test areas. The weights appear to be based on qualification standards as well as regression coefficients, and the resulting 20-point qualification levels are interpreted uniformly. There is a profile booklet, and a qualification grid booklet for computing and recording the 24 job-area qualification scores. These tests are intended primarily for use in employment and placement, and in consequence the manual and notes are addressed mainly to business and industrial executives. Much of the material in them is frankly educational rather than strictly technical and descriptive. Professors of measurement may be surprised on reading them to discover how much of personnel testing theory can be presented in quite practical and non-technical language.

The Guilford-Zimmerman Aptitude Survey consists at present of the seven booklets of Form A. Separate answer sheets may be used with the tests of verbal comprehension, general reasoning, spatial visualization, and mechanical knowledge (the power tests), and if necessary with spatial orientation also, though this is not recommended. The speed tests, numerical operations and perceptual speed, must be booklet-marked. Form B is in preparation, and the authors intend to expand the battery eventually to some 20-odd tests. A profile chart gives C-scores (an eleven-point scale with mean 5 and standard deviation 2), T-scores (mean 50 and standard deviation 10), and centiles for college men and college women, based on the scores of approximately 2,700 men and 1,500 women, mostly Freshmen, at the University of Washington, Northwestern University, and the University of Southern California. Reliabilities, intercorrelations among the subtests, and validities for grades in several dozen college subjects are reported, along with a few correlations with occupational criteria, and studies verifying the factorial structure.

The revised General Aptitude Test Battery now has eight paper-and-pencil tests and two apparatus tests. Three of the original paper-and-pencil tests have been eliminated, and two of the original factors (aiming and motor speed) combined. With the exception of mark making, a pure speed test, booklet marked, there are two parallel answer-sheet marked forms of each of the paper-and-pencil tests. The statistical data reflect the resources of a government bureau. The basic norms were determined for the earlier edition on a sample of 4,000 employed persons, selected from more than 8,000 to be representative of the general working population in terms of occupation, age, and region, and consisting of approximately equal numbers of men and women. The new Form A was standardized by equating its standard scores to those of the original form on a sample 585 high school and junior college students who took both forms. The new Form B was standardized in the same manner against Form A on a sample of 412 high school juniors and seniors. Aptitude score norms are weighted-composite standard scores with mean 100 and standard deviation 20, with weights which adjust for differences in both raw-score standard deviations and regressions of the tests on the factors defining the aptitudes. Seven studies report reliabilities, three report intercorrelations of the tests, three report intercorrelations of the aptitudes, ten report correlations with college grades, and seven report correlations with other batteries and tests. Other studies report sex differences and increase of scores with age. Occupational aptitude patterns consist of minimum scores on subsets of three or four aptitudes. This is a multiple cut-off system. Minimum scores are given for 17 job families and five ungrouped occupations, based on 99 validity studies yielding significant results. A profile card and a card giving the minimum scores for job families and ungrouped occupations are supplied for the use of State employment offices. Unlike the other tests here considered, the General Aptitude Test Battery is not for sale, but wherever a legitimate testing need exists which can be met by it, the local office of the State Employment Service is usually glad to cooperate.

The six booklets of the Differential Aptitude Tests are re-usable, all responses being recorded on answer sheets, and there are two equivalent forms of each test. The norms are scores or score-ranges corresponding to 23 selected percentiles. They are given for each sex and each form for grades 8 to 12, and are based on the scores of over 47,000 pupils in over 100 school systems covering every region. A profile sheet is provided, with percentiles laid off on a normal-distribution scale, and a T-score scale in equal units beside it, with layout such that one inch represents the 1 percent significance level for the difference between the scores on the two tests. The T-score scale has mean 50 and standard deviation 10. Several thousand correlations are reported between the tests and school grades, and some of these are summarized in charts showing the distributions of coefficients for the major course areas: English, history and social studies, mathematics, and science. A considerable number of the studies show correlations between scores and course grades received 6 months to 3.5 years later, and still others show correlations with college freshman grades. Some hundreds of correlations with achievement test scores are reported, along with percentile equivalents of average test scores of students in various college curricula and in a dozen-odd occupational groups. Reliabilities are reported for each test by grade and sex, together with two sets of re-test coefficients after three years. Mean within-grade intercorrelations among the tests are reported for boys and girls, and correlations are also reported with ten other aptitude tests and batteries, and with the scales of the Kuder Preference Record.

The Multiple Aptitude Tests consist of nine booklets. They may be used with or without separate answer sheets. T-score and percentile norms by test are given for each sex for grades 7 to 13, based on 11,000 cases from eight regions. The T-scores are normalized standard scores with mean 50 and standard deviation 10. Differential intelligence norms give score ranges for each test, at 15 selected percentiles, for children of high, average, and low intelligence, at junior and senior high school levels, by sex. A profile chart is provided on which the standard scores of the tests may be averaged to obtain the factor scores. Reliabilities of the tests and factors are reported by grade and sex, and intercorrelations among the tests by sex for grades 7 to 13 combined. Correlations between each test and 15 other tests are reported by sex, as are also correlations with 16 school subjects, and these latter data are charted to exhibit differential validity. Additional studies report correlations with college freshman grades in four subjects at one college and in five at another, the latter separately by sex. Factor analyses of the two intercorrelation matrices are reported, along with data on the validities of the factor scores for scholastic performance. Expectancy tables for school marks predicted by separate tests are given for 19 combinations of test and subject. No correlations with occupational criteria are reported.

The Flanagan Aptitude Classification Tests come in fourteen booklets. They are booklet marked, with carbon-back self-scoring grids, and each grid contains a small table for converting the raw scores to stanines. There are alternate forms (Form B) for inspection, coding, memory, scales, arithmetic, patterns, tables, and mechanics, and the author plans to extend the series with additional tests now in preparation. The stanines, with mean 5 and standard deviation 2, are based on the scores of a representative sample of 1,563 Pittsburgh high school seniors, and a supplementary table gives the percentiles for boys and girls corresponding to the stanines. An aptitude classification sheet replaces the usual profile chart. Subgroups of stanine scores are added, and a table on the aptitude classification sheet gives occupational stanines for 30 occupations and college aptitude. The selection of tests for each occupation was made on the basis of job analysis data and the validities of similar tests as reported in the literature. The occupational stanines were then computed from the distributions of the sums of stanines, for the selected groups of tests, using the data of the norms sample. They represent equal weighting of the tests in the subgroup, and do not include anything in the nature of cut-off scores. Intercorrelations among the tests are presented for the data of the norms sample, and reliability data based on smaller samples are reported for the tests and for nine representative occupational batteries. Validity coefficients are reported for seven occupations and three college curricula, based on merit promotions over a four-year period and grade-point averages respectively. More recent materials, not yet in the manuals, give the mean stanine scores on the tests for subgroups of the standardization group who said they were satisfied and successful in 23 occupations and 19 types of post-high-school specialized training courses. Percentile norms are also reported for the applicable occupational stanines for nine occupations and nine training groups, and on the basis of these studies minimum occupational stanines are in preparation.

I shall not attempt to review the scattered representation of tests from the third category which appear in Table 1. The most important of these is probably the Yale Educational Aptitude Test, the data for which appear in Crawford and Burnham's *Forecasting College Achievement.*

Aside from the data just cited, and those in Table 1, how do the various batteries compare with one another? There is certainly no single answer to this question. All of these batteries are modern in the best sense of the term, and well constructed. Dollar costs aside, the user will get from any one of them just about as much effective measurement as he pays for in testing time. He can safely select whichever battery measures most nearly what he wants to measure in the time available to his subjects.

A few general observations, however, may be in order. It seems to me that there is little value in striving for almost-pure factor scores. If this results in the same test appearing in two or more factor composites, and the factor scores are then used to predict external criteria, this test will receive undue weight in the larger predictor composite if the latter is based on equal weighting of the factor scores. If the weights for the factor scores are regression weights, this test will increase spuriously the correlation between the two factor scores to which it contributes, thus lowering their regression weights. The net result may be that the test itself is properly weighted, but the other tests in the two or more factor composites will be underweighted. This criticism applies to the General Aptitude Test Battery and the Multiple Aptitude Tests.

If almost-pure factor scores are derived from *different* combinations of tests, the several tests measuring each factor will correlate highly with one another. The proper objective of any multiple aptitude battery designed for service use is to include tests all of which have low intercorrelations but every one of which will be a valid predictor of at least one category of important criterion variables. Batteries yielding almost-pure factor scores reduce thereby the "factorial range" that might otherwise be obtained in the same testing time. The ACE Test for Primary Mental Abilities, Chicago Tests of Primary Mental Abilities, Holzinger-Crowder Uni-Factor Tests, and Multiple Aptitude Tests are subject to this criticism to perhaps a greater degree than the others.

The value of descriptive norms is a function of their representativeness for some defined group rather than merely of the number of cases on which they are based. Here all other authors would do well to study the methods by which the United States Employment Service arrived at norms representative of the general working population. They *may* be as lucky as was the USES with its preliminary norms based on the first 519 cases which came to hand, but again they may not! Grade norms, age norms, and percentile norms for groups of specified age and sex or specified grade and sex should mean what they say, and samples carefully stratified on all other major determiners of variability are necessary to attain this goal. For some purposes one can question the need for descriptive norms. For industrial tests the basic norms can be arbitrary. If the object is merely to derive a system which will equate the scores on the tests, a pick-up sample such as Flanagan's is entirely adequate. For such tests the real need is for representative samples of workers in the various occupations and occupational groups. No such samples have ever been found outside the armed services, so far as I am aware, and the task is certainly Herculean.

The reported reliability coefficients are of little value in comparing one battery with another. Most of them are of the split-half variety, and a few more have been computed by one of the Kuder-Richardson formulas. All such coefficients are spurious to greater or less degree with timed tests, the amount of spuriousness depending on the severity of the time limits in each case. A few authors have even reported such coefficients for pure speed tests. The numerical value of a test-retest or alternate-form reliability is in part a function of the time interval separating the two test sessions. Coefficients of this type are reported for intervals ranging from consecutive administration at one test session to three or four years. Every reliability coefficient varies with the range of talent of the examinees. The ranges reported cover grade groups, age groups, and occupation groups.

Finally, we have with us still in occupational testing, the argument of weighted composites *versus* multiple cut-off scores, with, among others, the USES on one side and Flanagan on the other. Far be it from me to try to settle this argument!

Table 1 Probable Factorial Compositions

FACTORS	
V	verbal knowledge
F	fluency (verbal, ideational, expression)
S	space (incl. orientation and visualization)
D	deduction (syllogisms)
R	reasoning (incl. induction)
M	mechanical knowledge
IM	immediate memory
N	number facility
A	aiming
MS	motor speed
P	perceptual speed (incl. clerical speed, form perception, and symbol discrimination)
CS	closure speed (figure unknown)
CF	closure flexibility (figure given)

OTHER ABILITIES	
SP	spelling
L	language usage
VD	visual discrimination
J	judgment
MA	mathematical achievement
SI	science information

APPARATUS FACTORS (not here considered)
finger dexterity (GATB)
manual dexterity (GATB)
motor (Factored Apt. Series)

Explanation of "Remarks": 1) Easy arithmetic computation is mostly N; hard is mostly R. 2) Cross-out: 4 or 5 elements; one which does not "belong" to be identified. 3) Identical forms: figure given, same figure to be identified among others differing only slightly, but still easily discriminated. 4) Surface development: pattern and object (as of sheet-metal).

I. ACE Test for Primary Mental Abilities (SRA)

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Completion	36	[1]	V,R	Word from definition
2. Figures	60(?)[2]		S	Slide vs. turn-over
3. Verbal enumeration	870		P,V	Pick words of given class from list
4. Letter grouping	30		R,V	Cross-out
5. Addition	120		N,R	Fairly easy
6. Arithmetic	20		R	Problems
[3]				
7. Same or opposite	100		V	
8. Multiplication	150		N	Easy
9. Number series	30		R	
10. Cards	40(?)[2]		S	Slide vs. turn-over
11. Number patterns	30		R	Incomplete matrices
12. Initials	25		IM	Initials-surname
[3]				
13. Identical forms	60		P	
14. Marks	20		R	Position-series completion
15. Mechanical movements	44		M,S,R	
16. Word-number	20		IM	
[Three booklets]	1655(?)			

Factor scores: data not available to writer.

II. Chicago Tests of Primary Mental Abilities, Ages 11-17 (SRA)

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Addition	70	6	N,R	Fairly easy
2. Multiplication	70	5	N,R	Fairly easy
3. Vocabulary	50	4	V	
4. Completion	45	6	V,F	Definition given; supply word
5. Figures	20(54)[2]	5	S	Slide vs. turn-over
6. Cards	20(54)[2]	5	S	
7. First letters	80	5	F	Write words or four-letter words with given first letter
8. Four-letter words	60	4	F	
9. Letter series	30	6	R	Series completion
10. Letter grouping	30	4	R	Cross-out
11. First names	20	8	IM	Write first, given last
[One booklet]	495(563)[2]	58		

Factor scores: Number (1+2), Verbal meaning (3+4), Space (5+6), Word fluency (7+8), Reasoning (9+10), Memory (11).

III. SRA Primary Mental Abilities, Ages 11-17.

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Verbal-meaning	50	4	V	Vocabulary
2. Space	20(54)[2]	5	S	Slide vs. turn-over
3. Reasoning	30	6	R	Letter series
4. Number	70	6	N,R	Addition, fairly easy
5. Word-fluency	70	5	F	First letter given, write words
[One booklet]	240(274)[2]	26		

Factor scores: same as test scores. Total score: V+S+2R+2N+F.

---

[1] These data from test booklet only; writer did not have manual.

[2] Maximum score, where different from number of items as in multiple-answer items, in parentheses.

[3] Horizontal line indicates end of multi-test booklet.

IV. SRA PRIMARY MENTAL ABILITIES, AGES 7-11.

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Words	36	8	V	Vocabulary
2. Pictures	37	8	V	Oral vocabulary
3. Space	27	7	S	Paper form board
4. Word-grouping	27	6	R,V	Cross-out
5. Figure-grouping	27	8	R	Cross-out
6. Perception	50	5	P	Identical forms
7. Number	52	5	N,R	Fairly easy
[One booklet]	256	47		

FACTOR scores: Verbal (1+2), Space (3), Reasoning (4+5), Perceptual speed (6), Number (7).

V. SRA PRIMARY MENTAL ABILITIES, AGES 5-7.

1. Verbal-meaning	49	*	V	Oral vocabulary
2. Perceptual-speed	30	1.5	P	Identical forms
3. Quantitative	27	*	N,R	Counting and arithmetic
4. Motor	80	1	A,MS	
5. Space	24	*	S,R	Paper form board
[One booklet]	210	*Not timed		

FACTOR scores: Same as test scores.

VI. HOLZINGER-CROWDER UNI-FACTOR TESTS (WORLD BOOK)

1. Word meaning	45	5	V	Vocabulary
2. Odd words	45	4.5	V	Synonyms
3. Boots	70	2.5	S	Slide *vs.* turn-over
4. Hatchets	70	2.5	S	
5. Mixed arithmetic	60	3	N	Easy
6. Remainders	60	3	N	After easy division
7. Mixed series	40	7	R	Letter and number series
8. Figure changes	40	7	R,D	Figure analogies
9. Teams	30	6	D,V	Syllogisms
[One booklet]	460	40.5		

FACTOR scores: Verbal (1+2), Spatial (3+4), Numerical (5+6), Reasoning (7+8+9); also Scholastic aptitude: 5(1+2)+(5+6)+3(7+8+9).

VII. FACTORED APTITUDE SERIES (INDUST. PSY.).

1. Office terms	54	5	V	Technical vocabulary
2. Sales terms	54	5	V	Technical vocabulary
3. Scientific terms[4]				
4. Mechanical terms[4]				
5. Tools	48	5	M	What goes with what
6. Judgment	54	5	R	Mixed series completion and cross-out
7. Differences[4]				
8. Numbers	54	5	R,N	Easy to fairly hard
9. Perception	54	5	P	Name and number checking
10. Precision	48	5	P	Identical forms
11. Fluency	184	6	F	Prefixes, suffixes, jobs, office equipment (write words)
12. Memory	36	2+3	IM	Names and pictures
13. Dimension	48	5	S	Pick left-right reversed picture
14. Parts	48	5	S,R	Paper form board
15. Blocks	32	5	S,N	Block counting (AGCT)
16. Dexterity	90+120+180	1+1+1	A,MS	Trace, check, dot
[Separate booklets]				

---

[4] These tests not available to writer for examination.

VIII. Guilford-Zimmerman Aptitude Survey (Sheridan).

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Verbal comprehension	72	25	V	Vocabulary
2. General reasoning	27	35	R	Arith. reasoning
3. Numerical operations	132	8	N,P	Easy
4. Perceptual speed	72	5	P	Identical forms
5. Spatial orientation	58	10	S,R	Boat-heading changes
6. Spatial visualization	60	30	S,R	3-dimensional rotations
7. Mechanical knowledge	55	30	M,V	20 picture; 35 verbal
[Seven booklets]	476	143		

IX. General Aptitude Test Battery (USES).

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Name comparison	150	6	P	
2. Computation	50	6	N,R	Fairly easy
3. Three-dimensional space	40	6	S	Surface development
4. Vocabulary	60	6	V	
————[3]				
5. Tool matching	49	5	P	Identical forms
6. Arithmetic reason	25	7	R	
7. Form matching	60	6	P	2 sets of scattered figures
————[3]				
8. Mark making	200	1	A,MS	
[Three booklets]	634	43		

Aptitude scores: Intelligence (3+4+6), Verbal (4), Numerical (2+6), Spatial (3), Form perception (5+7), Clerical perception (1), Motor coordination (8).

X. Differential Aptitude Tests (Psy. Corp.).

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Verbal reasoning	50	30	V,R,D	Verbal analogies
2. Numerical ability	40	30	R,N	Arith. comp., hard
3. Abstract reasoning	50	25	R	Figure progression
4. Space relations	40(100)[2]	30	S	Surface development
5. Mechanical reasoning	68	30	M	Mechanical comprehension
6. Clerical speed, accuracy	100	6	P	Identical symbols
7. Language usage				
I. Spelling	100	10	SP,V	Single word R-W
II. Sentences	50(95)[2]	25	L,V	Locate errors
[Seven booklets]	498(603)[2]	186		

---

[2] Maximum score, where different from number of items as in multiple-answer items, in ( ).

[3] Horizontal line indicates end of multi-test booklet.

XI. MULTIPLE APTITUDE TESTS (CAL. TEST BUR.).

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Word meaning	60	12	V	Vocabulary
2. Paragraph meaning	50	30	V,R	
3. Language usage	60(120)[2]	25	L,SP,V	Error location
4. Routine clerical facility	90(180)[2]	6.5 or 8[5]	P	Name and number checking
5. Arithmetic reasoning	35	30	R	
6. Arithmetic computation	35	22	R,N	Fairly hard
7. Applied science and mechanics	60	30	M,S,R	Mech. comp. and mech. movements
8. Spatial relations—two dimensions	25	8	S	Paper form board
9. Spatial relations—three dimensions	25	12	S,R	Surface development
[Nine booklets]	440(590)[2]	175.5		

FACTOR scores: Verbal comprehension (1+2+3), Perceptual speed (3+4), Numerical reasoning (5+6), Spatial visualization (7+8+9).

XII. FLANAGAN APTITUDE CLASSIFICATION TESTS (SRA).

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Inspection	40(155)[2]	6	P,VD	Identical forms
2. Coding	150	10	P,R,IM	Difficult
3. Memory	25	4	IM	Memory for code
4. Precision	252	8	A	Narrow path tracing
5. Assembly	20	12	R,S	3-dimen. paper form board
6. Scales	120	12	P,R,VD	Curve reading
7. Coordination	100	2′40″	A	Wide path tracing
8. Judgment and comprehension	24	35+	V,R	Paragraph reading
9. Arithmetic	125	10	N,R	Fairly easy
10. Patterns	30(60)[2]	20	CF,S,A	Copying designs
11. Components	40	20	CF	Like Gottschaldt
12. Tables	120	10	P,S	Table reading
13. Mechanics	20	20	M,R,S	Complex mech. movements
14. Expression	52(64)[2]	35+	L,V	Grammar and usage
[Fourteen booklets]	1118(1275)[2]	204+		

---

[2] Maximum score, where different from number of items as in multiple-answer items, in parentheses.

[5] 8 if separate answer sheet is used.

XIII. Yale Educational Aptitude Test Battery, Form B(ERB).

Test	No. of Items	Working Minutes	Probable Factors	Remarks
1. Paragraph reading	40	15-20	V,R	One wrong word
2. Word relations	65	15-20	V,R,D,L	Opposite of different part of speech
3. Synonyms	100	15-20	V	Vocabulary
4. Translation (Art. Lang.)	84	15-20	R,V,L	
5. Translation (Art. Lang.)	96	15-20	R,V,L	
6. Memory (Art. Lang.)	45	17-20	IM,R,V,L	Translate without key
7. Equations	70	15-20	R,MA	Algebra computations
8. Equations	62	15-20	R,MA	Problem: formula; functional change
9. Figures	41	15-20	R,MA	Geometry
10. Cubes	120	15-20	S,R	
11. Projections	20	10-12	R,S	
12. Composite figures	48	20-25	S,R	
[3]				
13. Word relations	40	12-17	V,D,R	Verbal analogies
14. Logical inference	39	13-18	D,R,V	Enthymemes
15. Interp. of expts.	40	16-21	R,D,V	
16. Number series	30	11-16	R	
17. Symbolic relationships	30	9-14	D,R	Symbolic syllogisms
18. Discovering principles	40	31-36	R,D,V	Functional relations tabulated
19. Mechan. movements	61	22-27	S,R,M	
20. Mechan. movements	38	20-25	R,M,S	
[Two booklets]	1109	316-411[6]		

Area (Test) scores: Verbal comprehension (1+2+3), Artificial Language (4+5+6), Mathematical aptitude (7+8+9), Spatial Relations (10+11+12), Verbal Reasoning (13+14+15), Quantitative Reasoning (16+17+18), Mechanical Ingenuity (19+20).

XIV. Aptitude Tests for Occupations (Cal. Test Bur.).

1. Personal-social aptitude	45	20	V,J,R	Paragraphs
2. Mechanical aptitudė	60	20	M,S,R	Mixed mech. and space items
3. Gen. sales aptitude	45	20	V,J,R	Paragraphs
4. Clerical routine aptitude	60	12	P,V,SP	Mixed checking, alphabet., spelling
5. Computational aptitude	45	15	R,N	Arith. comp. and estimation
6. Scientific aptitude	45	20	R,V,D,S	Mixed problems
[Six booklets]	300	107		

XV. Engineering and Physical Science Aptitude Test (Psy. Corp.).

1. Mathematics	25	15	R,MA	Algebra computations
2. Formulation	10	10	R,MA	Algebra problems: set up formula
3. Phys. Sci. comprehension	45	10	V,SI	Science information
4. Arithmetic reasoning	10	15	R	
5. Verbal comprehension	43	10	V	Vocabulary
6. Mechanical comprehen.	22	12	M	
[One booklet]	155	72		

XVI. Cleeton-Mason Vocational Aptitude Examination (McKnight).

1. General information	50	[1]	V	
2. Arithmetical reasoning	30		R	
3. Judgment in estimating	30		J,V,R	Est. No. of men in Navy in 1920, e.g.
4. Symbolic relationships	20		R,D	Figure analogies
5. Reading comprehension	25		V,R	Paragraph reading
6. Vocabulary	45		V	Word-definition matching
7. Interest	98			Interest and personality factors not here considered
8. Typical reactions	80			
[One booklet]	378			

---

[1] These data from test booklet only; writer did not have manual.

[3] Horizontal line indicates end of multi-test booklet.

[6] Depending on whether or not practice test was given first.

The Logic of and Assumptions Underlying Differential Testing

JOHN W. FRENCH

Let me start my discussion of differential testing by taking a typical practical problem in which differential testing applies. Suppose a student has the choice of entering fields A ,B, or C, where A, B ,and C are either academic courses or occupations. Let us assume that we have given suitable batteries of tests to previous groups of students and have followed up those students to obtain a quantitative measure of how successful or how happy the students became in pursuing A, B, and C. For this criterion measure of success or satisfaction even a dichotomy would be satisfactory.

Now we are asked by the student which field we would recommend for him: A, B, or C. Our choice of the statistical techniques to apply should depend on what the student wants to know. He probably doesn't know exactly what he wants to know. However, I think we can assume that he would like to enter the field in which he would be most happy and/or most successful. This means he needs information such as (1) his chance of obtaining a certain level of success or satisfaction in each field, and (2) his chance of obtaining greater success or satisfaction in one field as compared to that in any other field.

Let me compare two statistical techniques that are recommended for developing test batteries useful in guidance work; multiple discriminant analysis and multiple regression.

Those who recommend multiple discriminant analysis in this kind of guidance work attempt to answer the student's problem by showing him how much resemblance there is between his own test scores and the average test scores for people in fields A, B, or C. It is suggested to the student that he enter the field in which his colleagues would have test scores most closely resembling his own. If the criterion groups for fields A, B, and C were chosen from among successful people in their respective fields, it is expected that the student will also be successful when associated with the group that he most closely resembles. How successful? What chance does he have of not being successful? Is he likely to be more successful in one field than in another? Multiple discriminant analysis doesn't answer these questions. It is an excellent technique for detecting membership in a group, for handling the very elusive problems of classification based on qualitative differences. But it does not answer the question: "How well will I do if I take a job as a dog catcher?" Although discriminant analysis cannot answer this kind of question, it does have a place in guidance work. It is probably the best available method in cases where criterion scores are unavailable or so restricted in range that multiple regression would give only a distorted picture. I will discuss this limitation of multiple regression later.

Validity coefficients rather than score patterns are the stock-in-trade for those who have satisfactory criterion scores available to them and who want to give what seems to me to be the direct answer to the student's problem. This is the multiple regression method. It provides predictions which indicate to the student his chances for attaining a given amount of success in A, B, and C, and differential predictions which indicate his chances for being more successful in one field than in another.

Let us look at the data in an actual case so that we can compare a counselor's advice based on multiple discriminant analysis with a counselor's advice based on multiple regression. Tables 1 and 2 on the handout present small portions from each of two larger tables. The rows in the two tables represent four aptitude scores: Perceptual Speed (this is mainly speed in finding given symbols in a mass of distracting material), Mechanical Knowledge (this is a knowledge of mechanical techniques and equipment), Carelessness (this is the number of errors made on speeded tests; a high score indicates many careless errors), and Speed of Judgment (this is the number of simple choices made within a short time limit; no attention is paid to the correctness of the subject's judgments or to the nature of his preferences). The columns in the tables represent groups of vocational high school students who later became respectively office workers, beauty operators, carpenters, and mechanics. The first two groups are girls, the second two are boys. Table 1 gives the validity coefficients for vocational shop course grades. Blanks occur in the table where the coefficients were non-significant. Table 2 gives the mean test scores for the four groups of students. For convenience of interpretation the means have been converted so that 50 is the general mean of all groups and 10 is the standard deviation.

For the office worker group, Perceptual Speed and Speed of Judgment look good from the standpoint of the validity coefficients. Therefore, multiple regression would choose office workers who had high scores on these two aptitudes. Future office workers also have the highest mean scores on these two factors. Therefore, multiple discriminant analysis would guide into office jobs girls who had high scores on Perceptual Speed and Speed of Judgment. Thus, here is a case where both multiple regression and multiple discriminant analysis would select the same people for the job.

For mechanics the validity coefficients recommend high mechanical knowledge, carefulness (that is, there is a negative validity for number of careless errors), and slowness of judgment (there is a negative validity for number of choices made). The means, on the other hand, show that the criterion group of mechanics had high mechanical knowledge, but they were the most careless of the four groups and were speedier of judgment than the carpenters. This is a situation where multiple regression would guide different boys into mechanics than would multiple discriminant analysis.

For beauticians and carpenters the two methods would also select somewhat different kinds of people.

Which method is the more suitable? Let me reply by asking a leading question. Do we want to encourage speedy, careless boys to go into mechanics just because mechanics are speedy and careless now, even though speed and carelessness correlate negatively with performance ratings?

I have tried to point out how two theoretical models for differential testing are related to the practical problem of counseling. The multiple regression techniques when made possible by the nature of the data seem to be more suitable at least in view of the kind of discussion I have been advancing. Let me now turn to a discussion of some of the theory bearing upon the accuracy and the limitations of predicting amount of success by the multiple regression techniques.

There are two ways for measuring the effectiveness of differential testing that make pretty good sense to me. By inspecting the equations involved it is possible to understand what things need to be maximized or minimized to attain the most accurate discriminations.

Paul Horst has developed a number of general formulas in this area. William Mollenkopf[1] has worked out a formula for the validity of a battery in predicting a difference between two criteria, a and b. This formula is Formula 1 on the handout. R_{d*d} is the validity of the differential prediction, that is the correlation between d_*, the predicted difference, and d, the observed difference. Stars in this notation mean *predicted*. R_{a*a} is the validity of the battery for criterion a, and R_{b*b} is the validity for criterion b. r_{a*b*} is the correlation between the predicted criterion scores, and r_{ab} is the correlation between the observed criterion measures.

---

[1] MOLLENKOPF, W. G. Predicted differences and differences between predictions. *Psychometrika*, 1950, *15*, 409-417.

It is clear from the equation that the validities for the two criteria should be high. r_{ab}, the correlation between actual criteria, depends upon what particular criteria are involved and so is not in the experimenter's control. The critical point for Mollenkopf's equation is that the correlation between predictions should be as low as possible. Let me translate this demand of the equation into terms of direct interest to the constructer of the test battery. Let us suppose that each test in the battery had the same validity for criterion a as it had for criterion b. For example, suppose we are trying to discriminate between plumbing and carpentry. Perhaps a mechanical test has a high validity for both. Let's say a verbal test has a low validity for both. Then the same tests and same weights would be used to predict success in both plumbing and carpentry. The predictions for any one person would be exactly the same. $r_{a^*b^*}$ would be 1.00, and, according to Formula 1, the validity of differential prediction would be zero. On the other hand, if each test has a very different validity for plumbing from what it has for carpentry, the predictions for the two criteria will be made on the basis of different tests or very differently weighted tests. The correlations between predictions, $r_{a^*b^*}$, will be a minimum. That is, it is a critical requirement for each test to have different validities for the different criteria. This differential validity is more likely to occur if the tests in the battery are highly independent one from another. Use of pure-factor tests or factor scores is one way to heighten chances of reaching this goal. The validity coefficients in Table 1 on the handout indicate that here is an instance where some success was attained in finding for each test widely different validities for the different criteria.

It is perhaps wise to remind ourselves here not to lose sight of the fact that good general prediction is also useful in counseling. That is, the student not only wants to know in which job he will do best, but he also wants to know how well he is likely to do. One should, therefore, consider the inclusion of some highly valid tests of mixed factorial content. There is a real danger of losing high general prediction when one is trying too hard to get good differential prediction.

Another way of judging the effectiveness of differential prediction that makes good sense to me was first described by T. L. Kelley[2] and later developed by Segel[3] and by Bennett and Doppelt[4]. Suppose two persons stand at exactly the same level on some aptitude. When these two people are tested for this aptitude by fallible tests, there will be a difference between the scores they receive. If the testing is done repeatedly, a distribution of differences will evolve. This distribution of differences may be said to be entirely attributable to chance, since there is actually no difference between the aptitude levels of the two people. In the case where a real difference in aptitude level does exist, the observed differences in scores will be greater; they will be partly attributable to chance and partly a reflection of the real difference in aptitude level. The effectiveness of differential testing can be stated in terms of the proportion of observed differences that are not attributable to chance. If the two variables in question are highly related, the real differences will be small. Therefore, the proportion that is *not* accounted for by chance will be low. If the two variables are relatively independent, the real differences will be large. If the tests are highly reliable, the chance differences will be small, and the proportion *not* accounted for by chance will be high.

---

[2] Kelley, T. L. A new method for determining the significance of differences in intelligence and achievement test scores. *J. educ. Psychol.*, 1923, *14*, 321-333.

[3] Segel, D. *Differential Diagnosis*. Baltimore: Warwick and York, 1934.

[4] Bennett, G. E., and Doppelt, J. E. The evaluation of pairs of tests for guidance use. *Educ. psychol. Measmt.*, 1948, *8*, 319-325.

For computing the proportion *not* accounted for by chance, Bennett and Doppelt presented an easy-to-use nomograph. Kelley presented a table yielding the desired proportion when entered by Formula 2 on the handout. In this value the numerator gives the standard error of differences caused by the unreliability of the tests, and the denominator gives the over-all standard error of differences found between test scores. In the equation R_{1I} and R_{2II} are the reliabilities of the tests, and R_{12} is the correlation between the test scores.

While this formula was worked out for pairs of individual tests, there is no reason why it cannot be applied to pairs of test batteries. When we are interested in prediction, the batteries used for two criteria will usually overlap, because one or more of the tests are likely to be valid for both criteria. The correlations between the predictions for the two criteria are likely, therefore, to be high. The correlation between predictions is analogous to the R_{12} in the formula. The formula shows that it is critical to keep this correlation down. This can only be done by having relatively independent tests weighted as differently as possible in the prediction equations. This means that here again each test must have widely different validities for different criteria.

There is one very disturbing matter that seems fitting to discuss in connection with the foregoing remarks about highly differential validities and about the choice between multiple regression and multiple discriminant analysis. It is something that tends to befuddle the multiple regression approach to differential prediction.

Let's say we are trying to predict success as a mechanic. In view of the correlations appearing on the handout the regression equation for this prediction will include a considerable weighting of Mechanical Knowledge and a smaller negative weighting of Carelessness and Speed of Judgment (or positive weighting of Carefulness and Slowness of Judgment). Now let's suppose that a hypothetical factor X was also, for some obvious psychological reason, absolutely essential for mechanics, so essential that all mechanics need it in a high degree. This factor X might be some such thing as a willingness to get all messed up with dirty grease. The range of scores on factor X would be at a high level and very restricted in extent. This would make the observed validity coefficient for factor X low, perhaps so low that factor X would not enter into the prediction equation for mechanics at all. Suppose we used only the factors with high validities to make our predictions. Then we might predict that a certain student would do well as a mechanic, because he is high on Mechanical Knowledge and low on Carelessness and Speed of Judgment. Nevertheless, he might fail completely, because he lacked factor X.

This kind of error can be avoided in either of two ways. One way would be to apply a special cutting score in cases of variables like factor X. For example, a student would be given no prediction for success as a mechanic unless his factor X score fell within the range which the criterion group of mechanics had for factor X. That is, unless the student is willing to get messed up with dirty grease, you don't predict his success as a mechanic at all. If his factor X score was in the proper range, his success in mechanics would then be properly predicted by the regression equation computed from uncorrected validity coefficients. For such individuals whose factor X scores were already known to be within this high range, the amount of factor X possessed by them might be sufficient for success as a mechanic, and therefore not important in predicting amount of success. That is why the low validity coefficient of factor X would be appropriate, provided factor X was used separately to eliminate those whose scores on it are low.

The GATB takes this matter into account through the rules it uses for selecting the "key aptitudes" upon which the qualification of individuals for jobs is based. Among these rules are the provisions that aptitudes should be considered as "key aptitudes" for a particular job if the mean score for people in that job is high relative to the mean score of the general population and if the standard deviation of the scores for people in that job is low relative to that for the general population. By selecting "key aptitudes" in this way, GATB is giving extra weight to the aptitudes which are thought to be so important to a job that their range of scores for people on the job is high and restricted. The added weight given to such aptitudes will quite properly tend to offset the lowering of the observed validity coefficient due to restriction of range.

Now let's examine again what we are really doing when we use a variable for guidance just because its mean is high for a particular criterion. Let's also examine what we are really doing when we correct for restriction of range. In the example I mentioned it turned out that mechanics have a high mean in carelessness even though the criterion values correlate negatively with carelessness. If we guide students into mechanics just because they resemble our criterion group of mechanics, we are assuming erroneously that it is good for mechanics to be careless. Let's say we have found that some people who tried to be mechanics but could not make the grade were low on factor X. This would show factor X to have positive validity even though validity coefficients may not have revealed it. Or perhaps there is some psychological or practical reason that makes it logically apparent that mechanics should be high on factor X. If either of these things is so, it would be reasonable to guide into the mechanical trades only those students who were high on factor X.

Now take the case where we do *not* have an independent study showing the validity of any aptitude with restricted range and do *not* have any particular psychological reason for being sure that high scores on any aptitude are necessary for mechanics. If restriction of range on any one aptitude is extreme, we must, as I mentioned before, limit our predictions based on that aptitude to persons whose scores fall within the restricted range. If, on the other hand, restriction of range is, say, no greater than 50 per cent, it is possible to use the known range for mechanics and the known range for the total population to correct the obtained validity for restriction of range. When the corrected validity coefficient is used, the aptitude with a restricted range of scores should take its proper weighting in the regression equation, and any student whether within the restricted range or not can be given a prediction as to the amount of success he could expect if he entered mechanics.

This is all very satisfactory if the regression is linear. However, if there are no mechanics with low scores on factor X, we will not be able to tell whether it is linear. The lower part of the scatter plot of factor X scores versus mechanics criterion values does not exist. Linearity in this lower part of the scatter plot cannot be proved, but must be assumed in order to extrapolate the regression line to accommodate students with low values of factor X. If restriction is not more than 50 per cent the assumption is probably not more dangerous than many of the assumptions we have to make in the field of testing. However, some accuracy of prediction is lost by having to extend the regression line out beyond the range which served to locate it experimentally. Not only do such predictions of scores suffer from the usual error variance of the distribution of actual scores above and below the regression line, but there is also error variance resulting from errors in the determination of the slope of the regression line. Such errors become increasingly serious as the predictor score recedes from the mean of the criterion group. Snedecor[5] gives the formula for this variance. This is Formula 3 on the handout. The separate error variances are additive. The "1" in the parentheses is the usual error variance around the regression line. "1/N" represents the error in locating the mean through which the regression line must pass, and "$X^2/\Sigma X^2$," represents the error variance caused by errors in the slope of the regression line.

How serious a reduction in the accuracy of prediction is this? If, for example, the range of a predictor is restricted 50 per cent because the criterion group consists of very high scoring people on the predictor, a few students asking for guidance could be as far as eight standard deviations from the mean of the criterion group. Although this would be extreme, let's find out what the accuracy of prediction would be. With 100 cases $X^2/\Sigma X^2$ would equal .64, 1/N would be .01. The error variance, then, would be 65 per cent higher than the error variance for cases near the mean. The standard error of the predictions would be 29 per cent higher. This is enough to be considered, but is not very serious even for extreme cases as long as restriction of range is not over 50 per cent and as long as there are a reasonable number of cases in the experiment.

Again and again it seems that there is not one best method for doing something. The method depends upon the practical purpose. If a student wants to know how well he will succeed if he goes into mechanics, you should tell him how much he resembles the typical mechanic only if that is all you are able to tell him. Otherwise tell him what he wants to know. Estimate his likelihood of attaining a given amount of success. If a predictor has a restricted range for some criterion, *don't* correct for restriction of range if you consider people outside the range to be unqualified anyway, but *do* correct for restriction of range if you want to get the best prediction for people outside the range. The statisticians and psychometricians offer us an impressive inventory of formulas from which to choose. However, this does not always make the choosing easy. For me, I think it's like being a little boy facing the horrendous problem of choosing exactly the right piece of candy from a great big box.

---

[5] SNEDECOR, G. W. *Statistical Methods.* Ames, Iowa: Iowa State College Press, 1946, p. 120.

Table 1. Validity of factor scores for job training criteria.

	Office Workers	Beau-ticians	Car-penters	Me-chanics
Perceptual Speed	46	—	—	—
Mechanical Knowledge	—	—	39	36
Carelessness	—	33	—	−27
Speed of Judgment	31	37	—	−23

Table 2. Mean factor scores for students who entered the four jobs.

	Office Workers	Beau-ticians	Car-penters	Me-chanics
Perceptual Speed	58	52	47	47
Mechanical Knowledge	39	39	55	58
Carelessness	48	50	48	51
Speed of Judgment	53	51	48	49

Formula 1. Mollenkopf's formula for the validity of the prediction of a difference.

$$R_{d \cdot d} = \frac{\sqrt{R^2_{a \cdot a} + R^2_{b \cdot b} - 2R_{a \cdot a}R_{b \cdot b}r_{a \cdot b \cdot}}}{\sqrt{2(1 - r_{ab})}}$$

Formula 2. Kelley's formula for a value used in obtaining the proportion of differences not accounted for by chance.

$$\frac{\sigma_{d.\sigma\omega}}{\sigma_d} = \frac{\sqrt{2 - R_{11} - R_{211}}}{\sqrt{2 - 2R_{12}}}$$

Formula 3. Snedecor's formula for the standard error of a prediction for predictor scores not close to the mean of the criterion group.

$$S_{\bar{y}} = \sigma^2_{y \cdot x}(1 + 1/n + X^2/\Sigma X^2)$$

GENERAL MEETING

Communication of Test Information

Helping Students Understand Test Information

JOHN W. GUSTAD

The past fifteen years have seen developments in most branches of science and technology which even their greatest apologists would have felt to be impossible. Psychology in general and testing in particular have been in the van of these developments. Testing is quite a bit bigger business than it was when Wolfle (22) rendered an accounting just under ten years ago. While comparatively few Americans will, in their lifetimes, encounter psychologists directly, vast numbers will encounter tests. This will occur in school or college, in the military, in industry, in hospitals, clinics, or prisons. The chances of an individual's avoiding testing are rapidly approaching his chances of avoiding finger-printing, having chest X-rays, or paying income taxes.

There are numerous highly verbal critics who see or profess to see in this movement portents of the brave new world or of 1984. Zealous advocates are equally sure that God's in His heaven and all will be right with the world as soon as testing is applied to all human relations enterprises. As usual, the truth probably lies between these poles. Many psychologists are deeply concerned that test construction has lagged behind the rapidly developing science and that a technical product is being marketed in the name of psychology which does not represent the best thinking available. There are undoubtedly good reasons for these and other concerns. Growth spurts often bring with them some loss of coordination.

One group from which we have heard comparatively little but whose reactions should concern us greatly is made up of the rapidly growing pool of people who have been tested. These consumers have opinions; they also have money and votes. Since we professionals do most of the writing, the ideas of the consumers have not been well represented in the literature. The situation is especially critical in counseling and clinical psychology, for here much of the process rests on the assumption that the client will be willing to make use of information about himself derived in part from tests.

The vision which Parsons (16) incorporated in his book nearly half a century ago is becoming dim. There are good reasons for this, because his simple, three step scheme was somewhat too simple. Nevertheless, the general notion that one should analyze the individual, analyze the job, and match the individual and the job can still serve a useful purpose. When Parsons wrote his book, methods for individual analysis were few in number and crude in character. Today, a glance at Buros' latest volume (5) might be taken by some as *prima facie* evidence that there were more than enough analytic methods available. I doubt that many of us would accept this verdict whole-heartedly. Still, among the thousands of tests available, there are some whose validities and reliabilities are respectable enough to make them useful.

When tests are used administratively, as in the military establishment or in industry, administrators must consider public relations. Most will recall the furor associated with the introduction of the Selective Service Qualification Test. In the counseling situation, where client rapport is even more critical, where the usefulness of tests is measured —or should be measured—in terms of the adequacy of the decisions made by the client, we encounter problems striking at the very core of our operation. The opinions of clients are not known with any degree of accuracy; among counselors and clinicians, the dissatisfaction, the *malaise*, the gnawing uncertainty are acute.

Why, one might ask, can one not interview a client with a vocational choice problem, assign a battery of tests, give him the scores, and then expect that he will act as appropriately as the situation allows? This *modus operandi* was—and perhaps still is in some quarters—in effect for a long time with, it should be noted, not entirely bad effects. Yet most of us share some of the acute dissatisfaction with this approach.

Our colleagues with the well thumbed volumes of Freud's collected works on their shelves have pointed out that such procedures ignore the facts of life regarding motivation, conscious and unconscious. People, even college sophomores, have motives. Worse, these motives are dynamic, whatever that means. Sometimes, clients will not do their best on our tests. Most tests presume the presence of the old college try. On personality and interest tests, clients will sometimes lie to us, to themselves, or to both. Even if they do not lie very much and if they do try to answer the items to the best of their abilities, they will often refuse to believe or to act on the results of the tests they have taken. Anyone who has ever tried to convince an aspiring pre-medic that he just does not have the ability to make it, especially if a favorite uncle once patted him on the head and told him he was a real smart boy, will know what I mean. Most perverse of all, many will finally acquiesce on the surface but will, once outside the counselor's office, go on doing the same old maladaptive things, be that trying to get into medical school with a tenth percentile ACE score or trying to get through engineering school with an equally low score on the Engineering and Physical Sciences Aptitude Test.

It seems to me that the problem may be considered from two major points of view. First, we might well examine the client and especially the task which our tests have set for him. Second, we might consider the techniques used by counselors and clinicians in trying to help the client complete his task successfully. The client's task is, to a considerable extent, determined by the psychologist, and I would like to turn first of all to this aspect of the problem. As scientists in more or less good standing, we share a passion for precision and accuracy. We sometimes share the feeling of Samuel Butler who said, "I do not mind lying, but I hate inaccuracy." The language of numbers is rather natural for us, and sometimes it is productive. Moreover, we have a passion for speaking scientifically, which often means that we cover our tracks with qualifications so extensive and intricate that even we are sometimes in doubt about what our colleagues really are saying. Useful and proper as the language of numbers and standard errors is, it is not the language of the clients with whom we deal. Yet the process goes inexorably on with us following the currents in our science and drifting farther and farther away from the consumers of our technology.

Binet set out to measure intelligence. Most people think they know what intelligence is. Before he got very far on the way, Binet had introduced a strange new concept: mental age. Stern, searching for a metric by means of which to express this characteristic, put mental age into a ratio with chronological age, multiplied the whole *melange* by 100, and came up with the I. Q. This has become after forty years a household term, but by now most of us doubt its value and for the most part leave it out of our test development enterprises. Yet notice how far from the client's universe of discourse the first widely used test got and in how short a time.

The same pattern may be seen in the development of personality tests. Woodworth set out to accomplish a fairly straight-forward task: to sort out neurotics. Most people have some idea about neurosis. At least, they think it is a bad thing that has something to do with the personality. Perhaps this is enough. But what has happened in the past thirty-five years? Introversion-extraversion tests were developed. By the time these terms were becoming dimly understood, dominance and submission tests were the thing. Current tests locate the client on continua such as psychasthenia, FC and CF, D and Dd, W, rathymia, K, F, anxiety, repression, etc. How productive these particular traits are is not at issue here. The point is that we have, in groping toward a better understanding of personality, departed a great distance from the language of the client. It is, of course, true that some of these tests are not meant for the client's perusal but only for the counselor's edification. Nevertheless, the problem remains in many instances.

Some years back, Paterson and his colleagues in the Employment Stabilization Research Institute attempted to extend the psychograph principle. The occupational ability profile, while something of a misnomer, nevertheless represented an attempt to make test scores meaningful to counselors and clients, to come to terms with the dictum about a picture and a thousand words. The usefulness of occupational ability profiles for the personnel man has been fairly well demonstrated. Considerably less has been said about the client's problem of trying to learn about himself from the inspection of such profiles. One of the few thorough treatments is that of Bennett, Seashore, and Wesman (1). Profiles are still very much with us, but the more expert the counselor or clinician, the more he sees or professes to see in the relationships among the points in the profile. Clearly, profile analysis as it is usually practiced is not for the college sophomore. Parenthetically, I am somewhat intrigued by the different treatment afforded to profiles of ability scores and those of personality or interest scores. In the latter case, the interpretations often border on the mystifying. The MMPI, Strong, and Rorschach seem especially vulnerable. Except for some attempts with the Wechsler, I know of few instances where people have become particularly "dynamic" with profiles of ability scores. This leads me to wonder whether we are missing the boat in interpreting ability profiles or whether the interpretations of personality and interest profiles represent rather stupefying metaphysical leaps. Only time—and good criteria—will tell.

There is another line of development which has perforce contributed to the present difficulties. Binet worked hard to measure a global trait, intelligence. Other test constructors followed suit with tests of neuroticism, adjustment, mechanical aptitude, etc. Increasingly, there has been a tendency to try to measure pure traits. This has arisen largely as a result of the developments in factor analysis. I happen to be among those who believe that this line of endeavor will in the long run pay off with better tests and better descriptions of human behavior. The problem with which I am concerned here, however, is the intelligibility of test scores to the client. I would like to repeat here a notion I first expressed several years ago (14), namely, that the difficulty of test interpretation is inversely related to the counselor's understanding of the trait measured and to its predictive significance. It seems unlikely that we should try to give all of our clients a short course in psychometric and factor theory so that they will understand our tests. This task is hard enough with graduate students.

It would be possible to go on at considerable length documenting the difficulties which a developing test technology and theory present to clients and counselors, but I hope that the point has been made adequately. We are in somewhat the same situation as the physicist who, when asked to describe a chair, quite accurately states that it is largely made up of empty space crisscrossed by wandering atoms. Such an answer is of comparatively little use to a person who wishes to know whether or not he should sit down and, if so, what the consequences will be. I am certainly not proposing that we return to the measurement of the old, complex, global traits like mechanical aptitude and general intelligence. I am, however, suggesting that we have created a considerable gap between the client and his language and our tests and their language. Parenthetically, and related to this same area, we might do well to consider the problem of validity. I sometimes wonder how much rapport we lose when a client, trying to decide between medicine and engineering, takes an inventory which asks him whether he would rather be a motorman or a conductor. We are all aware of the predictive validity of such items, but clients are not. Perhaps something more might be done following Gulliksen's distinction (12) between intrinsic and correlational validity.

Turning now to the other issue, the counselor's methods, there has been growing for the past several years the feeling that our methods of introducing testing in the first place and of interpreting tests in the second have something to do with the problems we face in getting tests and test results accepted and acted upon. The general tenor of the arguments presented by Rogers (17) is too well known to need repeating. Among those happier with the use of tests in counseling, Bordin and Bixler (4) proposed that the process of test selection be considered an integral part of counseling, not an intruding element. They went on to suggest that the identification with the process achieved by encouraging the client to participate was worth any difficulties it might create.

The subsequent work of Seaman (19) and Dressel and Matteson (7) provided some substantiation for the ideas expressed by Bordin and Bixler. Seaman was interested in whether, in a permissive situation, clients would select appropriate tests in sufficient number. He concluded that they tended to do so. Dressel and Matteson went farther to study the effects of such involvement in the choice process on some outcomes of counseling. They found that client participation was positively related to improved self-understanding and to greater feeling of security in the choice made but not to satisfaction with counseling. A study recently completed at Maryland bears on the same point; discussion of it will be postponed until later.

With respect to client participation in test interpretation, much the same situation obtains. Bixler and Bixler (3) proposed that such participation would have salutory effects on counseling. Several studies provide partial substantiation. Dressel and Matteson (8), in another study, concluded that students who participated most gained correspondingly in self-understanding, in security with respect to the choice made, and in satisfaction with counseling. Kamm and Wrenn (15) concluded that client acceptance of test information was best when several conditions were met: first, when the client and counselor were completely at ease; second, when the client took a positive attitude throughout counseling; third, when the client was ready to respond on the basis of the new information; fourth, when the information presented was directly related to the client's problem; fifth, when the information presented was not in conflict with the client's self-concept. Kamm and Wrenn seem to be describing non-defensive clients. These are certainly desirable, but the techniques for reducing defensiveness are somewhat difficult to isolate.

Taking a slightly different tack, Rogers (18) compared two methods of counseling, one of which encouraged client participation, the other of which did not. He found no differences between the groups handled by the two methods, but he did find that higher level intelligence and more active client participation in counseling were related to better outcomes.

Intrigued by some of the same problems, we recently completed a study (13) at Maryland, conducted under a contract with the Office of Naval Research, dealing with different methods of test introduction and test interpretation and their effects on client learning as a dependent variable. Very briefly, we selected three methods of introducing and selecting tests, four methods of interpreting test results. The dependent variable was a discrepancy index employing differences between self-ratings and tested positions. The discrepancy index was adjusted for initial accuracy so that clients who showed high accuracy on pre-counseling ratings would not thereby be penalized in post-counseling ratings. Test introduction methods varied from extremely permissive to quite directive. Test interpretation methods included the use of profiles, verbal descriptions without visual aids, and two methods employing the clients' initial ratings which were compared with test scores.

Neither the rows nor the columns, introduction and interpretation methods, were related differentially to the dependent variable. Equal changes were observed for all groups. Moreover, the interaction term between interpretation and introduction was not significant. These results are in close agreement with those reported by Singer and Stefflre (20).

In connection with the same research project, Tuma (21) undertook to study certain personality characteristics of pairs of clients and counselors as these might be related to the dependent variable. His research followed the general line laid down by Fiedler (9) (10) (11). He found some relationships existing which suggested that methods as such, taken apart from the personalities involved, are perhaps not the most fruitful variables for study. He found, for instance, significant differences in average gains among clients seen by different counselors and significant correlations between client-counselor similarity indices on selected personality traits and the dependent variable. These correlations were significant only for the ability variables. Dominance, social participation, and social presence were the variables with the highest correlations.

A point to be kept in mind in the above studies concerns the different kinds of dependent variables employed. Singer and Stefflre, Tuma, and I employed adjusted discrepancy indices. Correlations between initial and final self-ratings and test scores have been used (2) as well as unadjusted discrepancy indices. All of these, it must be remembered, are inly intermediate criteria, not ultimate ones. Dependent variables seem in general to vary in availability inversely with their importance. Dressel (6) has summed up the case very well in the following:

> . . . our real concern . . . is only in part with the here and now; the ultimate concern is with the years after completion of school. Lacking the means for expensive follow-ups, recognizing the difficulty in attribution to counseling its exact contribution, and having a natural impatience for immediate action, we turn to criteria such as grades, graduation, stay in school, stability, or satisfaction with choice of major. Such criteria are not always applicable to all individuals in the same way and their relation to ultimate goals is not clear. (p. 71)

If I may summarize and perhaps over-simplify in doing so, it appears that the solution to the problem of how to make test scores meaningful to clients lies imbedded in the interpersonal relationships obtaining in the counseling interview. Moreover, techniques as such are probably not the final question; rather, we must seek to find those techniques which can be applied by selected counselors to appropriate clients. This is a large order.

In the meantime, I would like to reiterate my earlier point, namely, that we turn some attention to bridging the gap between our tests and our clients. I am certainly not proposing any abandonment of the search for better and more meaningful traits, but tests used in counseling are to a considerable extent useful in direct proportion to their intelligibility and acceptability to the client. Both for this kind of enterprise as well as for the work to be done on devising and revising techniques we need criteria which are closer to the life situations in which decisions are made and acted on. Until we get these, our research must remain under the cloud of suspicion that clients simply learn how, during the process of counseling, to say things that will make the counselor happy.

Since I have spent my time talking about problems and areas of ignorance rather than laying down nice clean, simple, guaranteed rules for making test information meaningful, I am afraid that this may have sounded like that most pedestrian of all prose productions, the doctoral dissertation. Rather than closing, then, with a plea for further research, I will read a couplet of Alexander Pope's which seems to sum up as well as anything the job we have to do:

> Men must be taught as if you taught them not,
> And things unknown proposed as things forgot.

REFERENCES

1. Bennett, G., Seashore, H., and Wesman, A. *Counseling from profiles*. New York: The Psychological Corporation, 1951.
2. Berdie, R. Changes in self-ratings as a method of evaluating counseling. *J. couns. Psychol.*, 1954, 1, 49-54.
3. Bixler, R. and Bixler, V. Test interpretation in vocational counseling. *Educ. psychol. Measmt.*, 1946, 6, 145-155.
4. Bordin, E. and Bixler, R. Test selection: a process of counseling. *Educ. psychol. Measmt.*, 1946, 6, 361-374.
5. Buros, O. *The fourth mental measurements yearbook*. Highland Park, N. J.: Gryphon Press, 1953.
6. Dressel, P. Evaluation of counseling. In Berdie, R. (ed.) *Concepts and programs of counseling*. Minneapolis: University of Minnesota Press, 1951.
7. Dressel, P. and Matteson, R. The effect of client participation in test introduction. *J. consult. Psychol.*, 1949, 13, 82-94.
8. Dressel, P. and Matteson, R. The effect of client participation in test interpretation. *Educ. psychol. Measmt.*, 1950, 10, 693-706.
9. Fiedler, F. The concept of the ideal therapeutic relationship. *J. consult. Psychol.*, 1950, 14, 239-245.
10. Fiedler, F. Comparison of therapeutic relationships in psychoanalytic, non-directive, and Adlerian theory. *J. consult. Psychol.*, 1950, 14, 436-445.
11. Fiedler, F. Factor analysis of psychoanalytic, non-directive, and Adlerian therapeutic relationships. *J. consult. Psychol.*, 1951, 15, 32-38.
12. Gulliksen, H. Intrinsic validity. *Amer. Psychol.*, 1950, 5, 511-517.
13. Gustad, J. The effects of differing methods of test selection and interpretation on learning in the interview. Final report, Office of Naval Research Contract Nonr 1225 (00), 1955. Mimeographed.
14. Gustad, J. Test information and learning in the counseling process. *Educ. psychol. Measmt.*, 1951, 11, 788-795.
15. Kamm, R. and Wrenn, C. Client acceptance of self-information in counseling. *Educ. psychol. Measmt.*, 1950, 10, 32-42.
16. Parsons, F. *Choosing a vocation*. New York: Houghton Mifflin, 1909.
17. Rogers, C. Psychometric tests and client centered counseling. *Educ. psychol. Measmt.*, 1946, 6, 139-144.

18. Rogers, L. A comparison of two kinds of test interpretation interview. *J. couns. Psychol.*, 1954, 1, 224-231.

19. Seaman, J. A study of preliminary interviewing methods in vocational counseling. *J. consult. Psychol.*, 1948, 12, 321-330.

20. Singer, S. and Stefflre, B. Analysis of the self-estimate in the evaluation of counseling, *J. couns. Psychol.*, 1954, 1, 252-255.

21. Tuma, A. An exploration of certain methodological and client-counselor personality characteristics as determinants of learning in the counseling of college students. Unpublished Ph. D. dissertation, University of Maryland, 1955.

22. Wolfle, D. Testing is big business. *Amer. Psychol.*, 1947, 2, 26.

The Obligations of the Test User

ALEXANDER G. WESMAN

The conscientious publisher of psychological and educational tests occupies an unusual, if not unique, position. Like the manufacturer of scientific apparatus, he is engaged in the production of instruments to meet the needs of professional people. Like the book publisher, he faces the problems of printing, of editing, of working with authors and their idiosyncrasies, of copyrights. Unlike the manufacturer of scientific apparatus, who can assume that the physicist, chemist or medical doctor understands the apparatus that is purchased, the test publisher can make no similar assumption. And unlike the book publisher, who does not need to concern himself with *who* reads his books (except that it be as many as possible) the test publisher must be constantly and actively concerned with those who use his products, lest those products fall into improper hands.

Further to complicate matters, the ethical publisher, having restricted his market according to the dictates of his conscience, still finds himself with purchasers whose preparation for the use of the published materials varies from complete knowledge and considerable sophistication to little or no training and dismaying naiveté.

The dictates of his conscience are not the only moral force acting on the publisher of educational and psychological tests and techniques. In recent years, much time and thought have been devoted to the consideration of his obligations to the professions and to the general public. Committees on Test Standards have been appointed by the American Psychological Association, the American Educational Research Association and the National Council on Measurements Used in Education for the express purpose of formulating specifications for tests and test manuals. The codes which emerged as a result of their deliberations have been reported by these associations in two pamphlets, copies of which should be in the hands of every test user. They are, on the whole, very sound documents; one hopes that the moral pressure they try to exert will, in the long run, prove beneficial.

Additional pressure is also directed at publishers by Buros' *Mental Measurement Yearbooks*, by test review forms in textbooks such as Cronbach's or Thorndike and Hagen's, and the critical reviews which appear in professional journals. These influences are forces for the good. How effective they really are, is unfortunately a matter for dispute.

Just a year or two ago, at one of these ETS Conferences, Oscar Buros offered the exasperated judgment that tests and test manuals published in recent years are not as good as many of those published a quarter of a century earlier. I doubt that many of his colleagues would adopt a similarly extreme position. At the same time, there are those of us who are cynical enough to believe that the mere existence of recommendations and reviews does not *ipso facto* improve the quality of instruments offered to the test user.

Over the years, it is neither the publisher nor the critic who most effectively determines the quality of tests; rather it is the test user. Unless the test user knows what a good test is, and withholds support from those which fail to meet high standards, the recommendations enunciated by organizational committees will be worse than ineffective—they will be put to harmful use as just one more device for deluding the innocent. A statement such as "the author has considered the Technical Recommendations set forth by the APA, AERA, and NCMUE in preparing this manual" could provide an aura of respectability which a given manual may not deserve, and the uncritical might well be misled. There is no Good Housekeeping seal of approval in the field of test publication; there is no substitute for professionally competent and conscientious judgment on the part of the test user. Test publishers have important professional obligations; test users have parallel responsibilities.

Test publishers should refrain from making unsubstantiated claims for the validity of the tests they offer; they should distinguish between what they hope, and what has been demonstrated. Test users should also be able to distinguish between what is hoped and what has been demonstrated; they should reject exaggerated claims of merit despite the attractiveness of the manual's format or the eminence of the author. Validity is a matter of the content of the test and the situation in which it is used. It is not assured by either the renown of the writer or reputation of the package designer.

It is proper and desirable for researchers to try instruments in new applications. One hopes, of course, that in the original selection of tests to be tried, some reasonable hypotheses have guided the researcher in his choice; this is not always the case. In any event, the researcher can make more or less of a contribution by publishing his results. If his results are positive, they serve to alert others of new situations in which a test may be effective; if negative, other researchers may be spared the futile effort of duplicating the experiment.

However, if the user applies a test in a situation for which neither the author nor publisher intended it, a negative result should not be construed as adverse criticism of the test. It may more appropriately be announced as a failure of the researcher's hypotheses to stand up. It is ironic that publishers and authors should so often be blamed when tests won't do what they were never intended to do; when the only fair comment on a study is "why did the researcher *expect* the test to be useful under such conditions?" The publisher may properly be taken to task if his tests don't work when they should; the tests should not be criticized if they don't work in situations for which they were not intended nor recommended.

The summer issue of *Personnel Psychology* contains an example of this abuse. A group of graduate engineers was given a series of tests including *DAT Space Relations*, *Mechanical Comprehension BB* and *Otis Arithmetic Reasoning*. The authors of the article reporting this research express surprise that the tests failed to discriminate among the engineers. The proper occasion for surprise is that these tests were chosen for use in this situation in the first place! They are good tests for the populations and purposes for which they were intended—high school students or unselected adults. That tests published for these levels do not yield adequate distributions for a group which has had intensive academic and professional experience with mechanical forces and advanced mathematics is hardly noteworthy. If the *Miller Analogies Test*, or the *Minnesota Engineering Analogies Test*, or the *GRE Advanced Mathematics Test* was not discriminative, we might criticize the test; with the tests selected for this study, we can only question the wisdom of the researchers.

Test publishers are constantly engaged in amassing evidence concerning the validity of their instruments in various applications and with different kinds of subjects. Test users must recognize that unless they provide the subjects to be tested, the needed data cannot be accumulated. Few and far-between are the occasions when a test publisher has a captive group of subjects at his mercy. More typically he is wholly dependent on the cooperation of the school administrator, counselor or teacher. The user has the right, and the duty, to refuse to buy a test which lacks proper documentation; he is also under some obligation to accept his proportionate share of the burden of providing a situation in which evidence concerning the test may be gathered during its exploratory, standardization and validation phases. It should not be left to the cooperative minority to provide the necessary subjects; all schools which hope to profit by the existence of good instruments should participate in experimental programs on appropriate tests.

A similar point may be made with respect to already existing tests. The publisher who neglects to collect serviceable normative data for his tests is properly to be criticized. Is less criticism due the non-cooperators in the schools—and in industry, government and private practice—who have useful normative data in their files but do not make those data available to the publisher and, through him, to their colleagues? How many millions of test scores repose in dusty files, or have been destroyed, which could have augmented the norms in the hundreds of manuals now in print?

Publishers should not over-emphasize the role which their tests should play in the over-all evaluation of a student, employee or client. The user might well apply an equal sense of perspective. It is difficult to say which has done the testing movement more harm—the naive optimist or the equally naive pessimist. The optimist looks to tests to solve all his evaluation problems—in effect, he surrenders the responsibility of personal judgment in exchange for the luxury of having something else make his decisions; often, it is a "something else" which was uncritically chosen in the first place. He operates as a clerk rather than as a professional man.

The naive pessimist, on the other hand, casts his jaundiced eye on the acknowledged limitations inherent in even the best of our tests. Though he would probably not say so boldly, he rejects the tests, in effect, because they don't have perfect reliability or perfect validity. If, in spite of his protestation, tests are used in his school, he warns in doleful tones that the scores must not be used alone as a basis for evaluating the individual.

We have no quarrel with the principle that a single test score—or for that matter, a series of test scores—should not provide the sole basis for action of any kind. Publishers typically urge users to correlate the information obtained from tests with all other relevant information that can be obtained, including grades in school, anecdotal records, physical reports, social workers' reports and whatever else local facilities permit. Our quarrel is that the naive pessimist wears blinders.

It is true that tests are not perfectly reliable or valid; is perfect reliability or validity to be found in grades? in anecdotes? in teacher observations? It is likewise true that tests alone are insufficient evidence for total evaluation of the student. Are we, however, to be satisfied with the evidence we obtain from grades alone? from anecdotes alone? from social workers' visits alone? One wonders whether it is not a sincere compliment (though perhaps unintended) that tests are singled out for warning with regard to their use in isolation; could it be that test scores are the only kind of information which would be considered tempting enough for such use? Nothing is that good, of course—but it is interesting that no one ever warns us about the isolated use of anecdotes or teacher observations.

The publisher has the obligation of keeping abreast of new developments in educational and psychological principles and practice, and of building tests which will reflect those modern concepts. The user is equally obligated to understand these newer instruments and the ideas they represent. We are sometimes told by administrators that, while they approve of intelligence measures with differential scores, such instruments can't be used because their teachers (or counselors) are used to the simple, single IQ. Is this reasonable? These same teachers are expected to look at a cumulative record showing grades in a variety of subjects and extract meaningful information. Multi-score achievement batteries are the rule, almost without exception; yet the teachers have presumably learned to interpret results from these tests. Why, then, should teachers and counselors be accused of inability to learn to interpret several scores on differential aptitude or intelligence tests? The logical answer seems to be that they can learn—if the administration takes its own responsibility seriously enough to provide the opportunity and motivation for learning. Modern medicine requires the general practitioner to understand the properties of modern wonder drugs. Modern education requires modern testing embodying modern concepts —and a willingness on the part of educators to continue their own education.

The preparation of a manual which provides the necessary instructions for administration, scoring and interpretation is an obvious duty of the publisher. Following those instructions is a parallel responsibility of the user. Every one of us, I dare say, has seen impossible scores reported on answer sheets, in personnel files or on cumulative record cards. I recall, for example, a set of records from a New York City school which contained half a dozen or so IQs of 400 and over, twice that many in the 300's and as for IQs of 200 or so, they were quite routine. I recall also a high school testing in Nebraska in which all but three or four of the seniors scored above the ninety-fifth percentile (national norms) on a clerical speed test. As my daughter would say, "Somebody goofed!"

One hopes that no responsible person gave serious credence to such outlandish scores, though their presence in official records does make one wonder. More serious than these dramatic bits of nonsense are the thousands of less dramatic, and consequently less conspicuous, scores which seem possible enough but which are really incorrect reflections of the testee's ability—misleading information as a result of someone's failure to read and heed test manuals.

The list of users' responsibilities could be expanded almost indefinitely; the points selected above are illustrative rather than exhaustive. The whole matter can perhaps best be summarized in two sentences. The publisher should feel obligated to prepare instruments which earn the user's respect by being psychometrically sound, conceptually modern, and administratively and economically practical. The user is under an even stronger obligation to cooperate in the development of these instruments and to support those which deserve support—not only in terms of purchase but also in terms of intelligent application and interpretation. The best portrait painter in the world would be handicapped by a house-painter's four-inch brush; the finest artist's brush obtainable would create no masterpiece in the hands of the untutored.

How Basic Organization Influences Testing

DAVID H. DINGILIAN

I. *Introduction*

Both World War I and World War II had a marked influence on the use of tests by the entire nation as well as by education.

Without any intent to be historically accurate in detail, it might be safe to say that, broadly viewed, World War I and the period immediately following it seemed to give impetus to the use of tests for the purpose of measuring and appraising the individual. During and since World War II tests have been used on a much broader scope. The users seemed to have added to the purposes of measurement and appraisal the desire to assess and understand the individual by way of counseling and other techniques.

Within learned circles of professional test users there undoubtedly exists a cautious estimation of the values gained and progress made because of the use of scientific tools such as tests.

One cannot help but wonder, however, if there is a parallel attitude on the part of the legion of inexpert test users who were willy-nilly thrown into the business of using tests without any disciplined orientation or training concerning the possibilities and limitations of those tests.

Here is a vice-president of one of the country's largest financial houses who says, "You psychologists sure have a black record with our outfit!" Then he proceeds to name three persons who came, tested, collected rather fat fees, and departed. Not too much was left behind in the way of insights about tests, except the hindsight of the management which indicated that it had an over-abundance of confusion and considerable unresolved feelings about "psychologists."

Here is a teaching colleague who comes to this writer with the delightfully naive, but nevertheless disconcerting, question: "I sent Jim down to the counselor and he tested him. Why hasn't his behavior improved? He *was* tested, wasn't he?"

Whether it is justifiable or not, we must face up to the fact that too many people have the uneasy feeling that psychologists specializing in the area of tests have promised too much too fast and, thus far, delivered altogether too little. To keep this kind of feeling from compounding, we must find ways wherein business and educational institutions, state and local, can use tests under greater scientific auspices.

One way to cut down further uneasiness about the use of tests is to invent ways whereby psychologists may be more active in testing programs. Let us take one field as an example. So much of the activities of the broad area known as counseling and guidance emanates from the seminars, practicums and laboratories of the behavioral sciences, particularly the discipline of psychology. How tragic to make such a fine contribution, say, for example, to the vital realm of public education, and yet have so many professional psychologists not close enough to the pulse of the activities to which they have given impetus!

In the main, departments of education and educational psychology have shown more interest in, and rendered greater professional assistance to, the test users than have departments of psychology. Yet, ironically, the large bulk of the tests themselves came from the fertile genius of rigorously trained psychologists. The question has occurred to this writer many times as to why psychologists do not seem to want to follow through to help toward the proper incorporation of the tools of their discipline into the organizational structure of such areas as education and industry.

It took the American people a long time to accept the common school as a basic institution. It took another seventy-five years to make the modern American high school part of the pattern of free public education. Recently, signs indicate that the junior college will be the next widely accepted institution. As one begins to rejoice about such matters, he senses that so much is being invested in buildings and excellent curricula without an equally adequate investment in orienting personnel to scientific methods. It almost looks as though we are guilty of extending a sentimental invitation which says, "Come one, come all—we are the servants of the next generation and offer you free education."

There must be an end to the developing of expensive and highly specialized schools and curricula without clearly defined criteria as to who should study what, how long, and where. Entrance requirements to numerous educational institutions are still entirely too devoid of any careful appraisal of the talents and abilities of the applicants. Most of these institutions are cluttered up with pupil-personnel whose assets probably consist of good intentions, a willing staff of public servants, and the democracy of opportunity.

Perhaps it would be appropriate to repeat a question raised by Daniel Starch in his address to the Invitational Conference of 1954. Some may remember that he stated, "It is fair to say that more advance in scientific knowledge has occurred during the last fifty years than in all previous centuries combined." Then he asked, "Why is there so wide a gap between what we know and what we do with what we know, between knowledge and the wisdom of how to use that knowledge?"

My first point is a disconcerting one. As a person who has spent the last ten years attempting to bring the use of tests into greater and greater prominence in education, I have found myself altogether too lonely. Persons in such roles as director of guidance, or director of a counseling center, are caught between a vast public, which wants more scientific help, and educational administrations, which have not been sufficiently structured about the values of tests by high echelon organizations from the professional area of psychology.

As one views the current scene, other than the reasonably favorable environs of the college and university campuses, he cannot help but be seriously disturbed about what has happened as a result of separating the tests from the insights of the test makers. Those who distribute tests may be said to be doing a good job of "selling" the idea of the use of tests. But, no matter how noble their efforts, their motivations are still linked by the test users with promotion for profit rather than with zeal for scientific rigor. Whether this is true or not is secondary to the fact that such motives are imputed to them.

In contrast to the harum-scarum growth of the use of tests in most elementary and secondary school situations are the numerous excellent student personnel programs which have been developed in many colleges and universities. There, one has a sense of orderliness and a feeling of experiencing closure. There is a rationale present. An appreciation of the values and limitations of tests is constantly a part of the multifaceted services being rendered. The user of tests at the college level is also, in most instances, the renderer of the service or the supervisor of those whom he is training to render service. Sound organization is present. Expediency and improvization are the exceptions rather than the rule.

The college or university pattern seems to include the setting up of structure and basic organization prior to the rendering of services which involve the use of tests. Why such a pattern has not jelled in elementary and secondary schools is a question which a conference such as this might explore.

It seems to this writer that the matter of centralization versus decentralization, organizationally speaking, is a false dichotomy. When one views both types of approach he soon sees that they can both be good or bad. The main variable may reside in the answer to the question, "Is there a doctor in the house, a doctor of psychology, that is?" One who knows tests, has constructed them, can teach the amateurs how, as well as how not, to use them. Most important of all, one who is sympathetic as well as flexible about the kind of in-service education which he himself may need. It is equally important for him to understand properly the purposes and objectives of the organization or institution which has sought out his scientific skills.

II. *An Example of a Testing Program in Action*

Having indulged in quite a bit of "Monday morning quarterbacking" calls for some kind of a constructive suggestion. What does happen when staff members of an organization take time to work out together a philosophy, organizational structure, and patterns of rendering services? It occurs to us that perhaps an example of a specific organization and its step-by-step evolving of an approach to the use of tests within the framework of counseling service might be of some interest at this point.

This writer was for eight years director of a counseling center which started out as a unit for serving veterans only. Within three years it was expanded to include a nine-to-twelve-hour service to graduating seniors of sixteen to eighteen high schools. The rendering of such services created kinds of problem situations which could have caused the organization to move in the direction of operating, either from improvization and expediency, or from a base of theoretical constructs which gave the staff a feeling of scientific precision and discipline.

When fully staffed, the center had a total personnel of seventy-five. The number and classifications of the staff were: fourteen psychologists, one of whom was the senior psychologist; twenty-eight counselors who came from a background of education; one supervisor and four assistant supervisors with backgrounds of education; and twenty-eight clerks, one of whom was principal clerk.

For nearly five years, this staff served a total annual case load of eleven to twelve thousand persons.

Ways of functioning which were entirely new had to be worked out. The staff decided to give priority to the evolving of a philosophy and pattern of organization. The fact that no previous organization, and hence no precedents, confronted the group was viewed as both a hazard and an advantage. The staff agreed to schedule only that amount of case load which would permit sufficient time to formulate, during in-service meetings, the beginnings of a philosophy and organizational structure. This firm resolve not to permit the pressure of case load to sidetrack adequate joint planning proved to be the most fruitful decision in the history of the guidance center.

A. *Philosophy*

In order to define a philosophy to use as a frame of reference, the staff listed six basic tenets. It isn't within the scope of this paper to elaborate the details of these tenets. A summary of the main ideas can only give us clues as to the values of consensus-making.

First, the complexness, as well as the vastness, of modern knowledge caused the staff to feel a need for synthesis. They realized that they had to be both specialists and generalists.

A modern synthesist, they decided, would create a mosaic made up of the findings of the various disciplines which deal with man. He would build a dynamic frame of reference which would depict "man and culture" as process, as change, as interaction, and as a continuum. In this approach, the synthesist would move from a point of departure to a point of view which, when sufficiently strengthened and enriched, could become a way of life; a way of life which would have one main objective: to instrument the scientific method toward serving the needs of our democracy.

A *second* tenet had to do with the hypothesis that all behavior is caused behavior. The individual's early environmental and inter-personal relationships as molded within such frameworks as the family constellation and peer groups, have given him an approach to life which he is now living out. The point which the staff agreed to emphasize was the necessity for approaching clients with an attitude of not condemning or condoning behavior, but understanding it as *caused behavior*.

The *third* tenet was viewed as being rather closely tied in with the second. Studying the outgrowth of nearly fifty years of clinical work in psychopathology and other related disciplines, and certainly, the influence of Rogers and his associates, helped the staff to explore the assumption that the client has great capacity and wisdom to help himself. They concluded that the client, if given certain psychological conditions, could reorganize his "field of perception"; he could alter the way he sees the field as well as himself, and hence modify his own behavior.

The *fourth* tenet of the staff's philosophy had to do with tests and testing. They agreed to be very careful that their possibilities, as well as limitations, were constantly kept well defined. They experienced the numerous ways in which tests so easily lend themselves toward being used as crutches. Counselor X would catch himself riding *this* interest test. Psychologist Z would become emotionally identified with *that* capacity test or projective technique. For the entire staff, tests and testing had to become just one of a series of factors making up the configuration which they decided to call advisement.

Having cautioned themselves about pitfalls, they agreed to develop the most complete and up-to-date test library which could be had. More than 225 of the so-called best standardized tests were procured. The psychologists were ever on the alert for new tests, new norm data, new validations. Four main objectives in regard to tests and testing were agreed upon. They were: (a) to use tests as means, and never as ends; (b) to have on hand the most scientific tools in order that the diagnostic work of the staff could be considered adequate by the best scientists in the field; (c) to avoid the pitfall of assuming that proper diagnosis is all that is needed, that it ends all the needs of all the clients; and (d) to be constantly aware of the special implications for test-users of the work of allied fields.

In formulating the *fifth* tenet in their philosophy of guidance, the staff recognized the need for an adequate concept of change. This meant discarding any tendency to see the illusory safety of absolute truth as a "thing" to cling to, and substituting the more challenging scientific "method." Concepts came to be valued in terms of their utility rather than validity. As the group sought to avoid the rigidity which comes with unchanging values, they tried to be wary of such dichotomies as nature versus nurture, introvert versus extrovert, or good versus bad.

The *sixth* tenet stressed the importance of having adequately defined democratic aims and methods. The group agreed that democracy is an idea which is much broader than a mere political or economic doctrine. They found that where it had been instrumented, it had tended to become a way of life, both social and individual. They agreed that its crux lay in the phrase "informed participation." Those staff members who came to believe sincerely in democratic consensus-making, came to know intimately the fact that where there is no participation there comes into play a subtle and insidious form of behavior. It changes the climate so that persons in authority flirt with the idea of suppressing further participation. Persons under authority begin to do their job with the attitude, "I merely work here." There begins a passivity, an indifference. These, in turn, grow into a sort of game between those who control and those being controlled.

During the actual functioning of the guidance center, the staff came to appreciate more and more the way in which an adequate philosophy could affect the quality of services rendered to the client. Among themselves, it fostered stimulating and worthwhile discussions as to the merits and demerits of different tools. And, in regard to tests, their philosophy caused the users to stop, look, and listen.

B. *Organization*

Once the beginnings of a philosophy were outlined, the next step seemed obvious, namely, the implementation of this philosophy into specific organizational structure.

Following are a few of the decisions which were agreed upon by the entire staff:

(1) Planning and policy-making should include not only the administrative and supervisory personnel but also a representative elected on a rotating basis from the clerical, psychological and counseling personnel.

(2) This group would be known as the administrative council. Each member, irrespective of rank or classification, would have equal responsibility to place on the regular agenda items which he felt to be important.

(3) In addition to the regularly scheduled meetings of the administrative council, each Thursday afternoon would be set aside for staff meetings. One of these would be the monthly general staff meeting. The other Thursdays would be used for special meetings of clerks, counselors and psychologists.

(4) Responsibility for conducting the meetings would rest with the group concerned. (a) The general staff meetings would be the responsibility of the head supervisor or of any of the other six supervisory assistants on the Center's table of organization. (b) The other meetings would be the joint responsibility of the assistant supervisor in charge of in-service education and the elected representative of each of the three classifications of personnel.

(5) The nature of these meetings would be: (a) business items pertinent to each of the three groups; (b) in-service education appropriate for each of the three roles.

(6) If the above machinery for communication did not adequately facilitate the two desired values of maximum participation and consensus-making, the administrative council would effect the necessary changes.

Again, we must say that greater detail on organization, than that outlined above, cannot be gone into because of the limited purpose of this paper. It is certainly true that such a structure became the anvil on which was hammered out every item that some staff member wanted explored. The items ran the gamut from the uses and abuses of the morning and afternoon coffee breaks to the most prolonged and detailed discussions of the relative merits of this versus that test, as being the most valuable tool to measure interest, abilities, temperament or aptitudes.

C. *Operations*

Some one has said that philosophy will not butter your bread, but it will make it taste better. The staff at the guidance center came to appreciate this maxim greatly. The operational plan which emanated from that philosophy was built on four basic concepts. *First*, it would be the responsibility of administrators and supervisors, at all times, to facilitate and not frustrate the processes of rendering guidance services. *Second*, establishing a flow of well-defined and orderly steps was imperative, because the quality of the end product would be determined by the nature and quality of the different guidance experiences offered to the client. *Third*, the client would have to be involved to the maximum in each of the steps which he would experience. *Fourth*, the weakest or the strongest staff member, be he a counselor, clerk, psychologist, or administrator, would be strengthened by the process if the various steps were well-defined. The assumption was that there is an objective kind of discipline involved in knowing what operational patterns go into gear in regard to such activities as the interview, test prescription, use of occupational information, or tentative selection by the client of three to five objectives.

Let us have a look at the steps involved in advisement: (1) *Structuring*—This can be done on an individual or group basis, at the center by the intake supervisor, or at a school by one of the members of the team of counselors and psychologists. The idea is to define the steps of the service, to tell what it does not do as well as what it does, to impress upon the client the importance of his role in increasing or decreasing the effectiveness of the service and to see that he has a chance to ask questions. The client needs help in seeing the point of view of the staff regarding the gathering and sharing of factual information, in contrast to their reluctance to interpret, advise or prescribe.

(2) *Basic Testing:* An interest inventory, a mental capacity test and a personality test are administered to each client. As a single client he may be able to go from structuring directly to testing. As a member of a group in one of the schools, he might have a time interval of one to three days between structuring and basic testing.

(3) *First Interview:* The client discusses with a counselor the following data: (a) his personal-social background; (b) the results of the basic tests; and (c) the making of a tentative list of ten to twenty vocational objectives which are compatible with the client's background information and basic test data.

(4) *Aptitude Testing:* The client explores with one of the psychologists, his counselor, or both, the kind of aptitude testing which is indicated.

The counselor needs to be sure that the information compiled in the first interview is the basis from which clues are gathered for the kinds of aptitude testing which are agreed upon.

(5) *Second Interview:* The counselor and the client review the results of the aptitude test, and relate such information to other test data, background information, and the list of ten to twenty tentative objectives previously selected. If no further testing is indicated, the client selects three to five of the most important vocational choices which he and the counselor agree upon. These may all be from the list of ten to twenty, or one or two new ones may be added.

(6) The client may invest the minimum of a half-day in studying his three to five objectives in the center's occupational information library. There an occupational information specialist and clerical assistants make available to the client and the counselor the kinds of data relevant to the client's vocational choices.

(7) *Terminating Interview:* After finishing his study of the three to five choices, the client returns to his counselor to discuss those which seem to be the most realistic for him. Together with the counselor, he outlines a plan for (a) further education, (b) immediate employment, (c) a training program or (d) need for further counseling if none of the first three can be agreed upon.

(8) *Parent Interview:* Upon completion of the guidance experience the counselor discusses with the client, if he is a young student rather than an adult, the availability of time for a parent interview. If the client concurs, he is asked to sign an information consent sheet, so that the counselor may feel free to discuss data about him with his parents.

(9) *Re-evaluation:* The client is invited to come back for any future review of his plans, if and when such a step becomes a felt need.

In this brief statement we have attempted to communicate two ideas. The first one has expressed our dissatisfaction with what we see on the current scene in regard to the use of tests. It seems to us that the scientist has made excellent progress in the area of inventing new tools in the form of tests. He has not, however, stayed on the scene in order to hold to a minimum those inevitable abuses which he alone could have anticipated when he saw tests falling into the hands of inexpert users.

The rigorous standards for proper test usage which good psychologists advocate have not been adequately upheld, even in the universities where student personnel services are so close to the birth of research.

The second thing which we have attempted is a brief description of a counseling organization which came to use tests in as scientific a manner as it was able to devise. Our assumption was that in such an example we would find clues as to how basic organization influences the use of tests.

III. *Some Generalizations*

We are now ready to make a few *generalizations*. We label them as such with the hope that perhaps some of them might contain the seeds for eventual restatements in the form of hypotheses. These could be tested in future research, for example, on the nature of the relationships between organizational structure and test usage. They are not worthy of being called hypotheses at this point.

(1) The first generalization which occurs to us is that, in the light of what has been said in sections I and II, the title of this paper might well be changed from, "How Basic Organization Influences Testing," to, "How Individual Needs of Clients Influence Both Basic Organization and Testing." It is our conviction that the self-realization needs of human beings seeking help, when respected and listened to, facilitate the setting up of unique organizational structure and creative test usage.

(2) The following innovations, which grew out of the experience of the organization are used as an example and have significance for the client, the professional staff serving him, and for education.

These innovations are: (a) Clients must be given an opportunity to become aware of the possibilities and limitations of any guidance service which they seek. This makes for economy in relationships and budget. Such economy is documented by the fact that the screened but not returned (SNR) percentage of the Center remained between two and three during the first eight years of its existence. At one time this writer was told by the directors of three large university guidance centers that their percentage of SNR's ranged between eighteen and twenty-two. They did not use the screening interview as a structuring technique. A receptionist asked the client to fill out some forms and go directly to the first step of the service, which usually was testing.

(b) Clients must be given their test results in circumstances which would be agreed upon by experts as containing the maximum in the way of learning conditions.

(c) Test and other data must be treated as confidential and belonging to the client. This approach helps him to feel more responsible as an active collaborator in the process which is designed to help him.

(d) The relationships between the client and the professionals who are helping him must be built upon the assumption that he will be treated as a person, in most instances, capable of making his own decisions without pressure. The concept of consequences takes on real significance when he learns that his own decisions as to occupational choices will be accepted even though they may be unrealistic.

(e) In order to be permissive, accepting, and treat the client as a collaborator, the staff needs a continuous program of in-service education. The minimum by-products of such a program usually are: greater skill in counseling, competence in using insights from psychology, and facility in handling sound vocational information.

(f) As a result of systematic and well-defined steps, which both professional workers and the client must experience in the process of using tests and doing counseling, serious oversimplifications are reduced, if not entirely eliminated. For example, the staff worker finally gives up the idea that he must *tell* and accepts the idea that he must *help*. As to the client, he gives up such notions as, "I must do what the tests or the psychologist tells me to do." He comes to accept such ideas as, "There is no one single occupational career for each person." He begins to feel comfortable about the fact that he can expect reasonable success in any one of the three to five objectives which he has selected.

(g) The approaches to the use of tests and counseling techniques described here have resulted in a new kind of knowledge about adolescents as well as the types of occupational objectives which they select. These have valuable implications for education in general and modern curriculum building in particular.

IV. *Questions*

Let us close with a few questions.

(1) Should not one of the testing services develop a model, so that a group which starts out on a venture such as the one described here does not flounder? It seems to us that just test research on norms, reliability and validity is not enough.

(2) Should the A.P.A., A.A.A.S., or some similar organization with status and scientific know-how, provide field staff services to help orient top echelon people in education and industry regarding the use of tests? E.T.S. is already helping in this regard.

(3) What is being done to disseminate, throughout the country, information about promising practices concerning the use of tests which have proven their merits?

(4) Would it be pertinent to explore carefully the values of a practice, now used in several states, of contracting with professionally competent agencies for the rendering of diagnostic testing services to be followed up with individual and/or group interpretation of results to pupils, parents and staff, as well as evaluation of the worth of such contracted services?

Let us close with the thought that, in the final analysis, the quality of any testing or counseling program is dependent upon the nature of the human relationships among those involved in the rendering and the receiving of the services.

If the relationships are built upon a foundation of mutual trust, acceptance, understanding and cooperation, the end-product of the program will be adequate and effective.

A good service is made up of more than testing; it is more than counseling or the giving of occupational or educational information. It is all these things plus a point of view and a way of life.

Mechanics, tools and techniques must remain means. The end must be a warm, contagious and humane program for all who honor us by saying, "I need help." We can do no less than to strive to achieve such a program.

The Psychologist and Society

MORRIS S. VITELES

The title of this address, *The Psychologist and Society*, is sufficiently broad in scope to permit a variety of approaches, at different levels of discourse, to the discussion of the impact of psychology upon society. The approach adopted for purposes of this meeting will appear as the discussion proceeds. However, in order to avoid initial misunderstanding, I might refer to a few aspects of the situation which I do not propose to discuss, even though their consideration could well fit into a conference on testing.

Specifically, for example, I shall not talk about tests and testing as such, although I might well enjoy the opportunity to talk about the effects upon the individual and upon society of testing programs which range from the diagnosis of feeble-mindedness, about which the psychologist knows something, to the use of inadequately validated clinical methods in the identification of executive talent, concerning the nature of which little is known.

I could show equal feeling in talking about several consequences which ensue from success in undermining confidence in teachers and in schools through a publication in which partial data and quotations are used to give credence to conclusions which are even contrary to those reached by the research investigators themselves. There is no doubt that the psychologist *can* and *does* render a great disservice to society when he employs such tactics, characteristic of the propagandist and of journalistic irresponsibility, in dealing with even such a limited area of human activity as the acquisition of basic skills.

In spite of the temptation to talk of such relatively simple impacts of the psychologist upon society, I have chosen, instead, to devote this talk to the more complex and higher levels of discourse which come to the fore when the psychologist undertakes to remake society itself.[*]

---

[*] The succeeding portion of this paper is taken from an address delivered at the closing session of the 12th International Congress of Applied Psychology (London, 1955) and published under the title of *The new utopia in science*. *Science*, 1955, **122**, No. 3181, 1167-71. Acknowledgment is made to *Science* for permission to reprint this material.

Throughout the history of civilization man has been intrigued by the possibility of remaking this unsatisfactory world into a better one—one formed in the image of his personal perceptions, aspirations, and values. In saying this, I do not have in mind the broad conceptualizations of philosophers and religious leaders, such as the Ten Commandments of Moses, the Golden Rule of Jesus, the Five Relationships of Confucius, the Four Noble Truths and the Noble Eightfold Path of Buddha, or other ideal standards of conduct that have exercised tremendous influence in a variety of very different cultures. On the contrary, I am referring to detailed plans for reordering the formal organization of the community, for spelling out the structure, the details of daily life, and the specific patterns of individual form and conduct. Exemplified early in Plato's *Republic*, such projects have, through the writings of Sir Thomas More, made the word *utopia* a commonplace conception in the languages of the world.

Plans for creating similar seats of "ideally perfect society and political life" (1) have come from a variety of sources. Literary men such as Samuel Butler (2) in England, Edward Bellamy (3) in the United States, and in a sense Cyrano de Bergerac (4) in France—to mention only a few—found means for describing the inadequacies of civilizations known to them and fertile outlets for their imaginations in the design of fairer worlds—in the pursuit of the perfect way of life, or in the words of Matthew Arnold, of "sweetness and light" (5) as a way of life.

Until recently the architects of utopia have, perforce, found it necessary to accept man as he is and to satisfy themselves with manipulating his environment and his institutional relationships—primarily economic and political—as a way of remolding the world and, as the great son of the Persian tentmaker wrote, bringing it "nearer to the Heart's desire" (6). It will be recalled that Rousseau, in fact, took the position that man himself—*natural man*—is a noble creature, corrupted only by the artificial and degrading civilization imposed on him (7). The utopias of Rousseau and of his literary disciples such as Chateaubriand (8), were thus quite consistently characterized by a rejection of the artificial trappings of so-called civilized life and a return to primitive forms of existence.

Utopian Engineering by the Psychologist

Today, by contrast, the creators of a "brave new world" undertake their task with avowed capacity actually to remake man himself and thereby to achieve the states of *inner* and *outer* perfection which, in the past, were promised only in the afterlife. As illustrated in the satirical novel by Aldous Huxley (9), biology furnishes the mechanism for modifying inherent and supposedly inflexible characteristics of the individual by manipulation of the embryo itself; physiology and psychology provide the tools for early and complete conditioning of the individual to a man-made world of perfected order.

The application of such psychological tools for this purpose finds even more concrete expression in the creation of *Walden Two* (10), a new utopia designed by the outstanding American psychologist Burrhus F. Skinner. Here, with unbounded faith in the capacity of a science of human behavior to change such behavior, Skinner subordinates "natural man" to the socially adaptive and conforming influences of scientific methodology.

Skinner's approach to a new utopia is epitomized in the answer given by the founder of *Walden Two* to a question bearing on the failure of earlier attempts to establish perfected centers of community living. The crucial fault, he points out, was the absence of *psychological management*. "The cultural pattern was usually a matter of revealed truths and not open to experimental modification—except when conspicuously unsuccessful. The community wasn't set up as a real experiment, but to put certain principles into practice. These principles, when not revealed by God, flowed from a philosophy of perfectionism. Generally, the plan was to get away from government and to allow the natural virtue of man to assert itself. What more," adds Frazier, the fictional protagonist of the new utopia, "can you ask for as an explanation of failure?" (10, p. 129).

Beliefs underlying this approach find expression in Skinner's scholarly writings, particularly in his book *Science and Human Behavior* (11). It is here that Skinner commits himself to the view that the deliberate manipulation or control of cultural practices and human behavior is a necessary feature of any civilization and the road to progress toward a better way of life. It is here also that he formulates *survival* as a criterion in evaluating control practices. Likewise, the crucial role assigned to a science of human behavior in relation to controlled cultural change is made apparent in this text. "We have," he writes, "no reason to believe that any cultural practice is always right or wrong according to some principle or value regardless of the circumstances . . . Science," he adds, "helps us in deciding between alternative courses of action by making past consequences effective in determining future conduct. . . . The formalized experience of science, added to the practical experience of the individual in a complex set of circumstances, offers the best basis for effective action" (11, p. 436).

It is noted by Skinner that experimentation involving control of cultural practices may yield findings that are distasteful to Western thought, which has emphasized the importance and dignity of the individual and the philosophy—accepted, according to Skinner, by many schools of psychotherapy—that "man is the master of his own fate" (12, pp. 44-68). "If," he concludes, "science does not confirm the assumptions of freedom, initiative, and responsibility in the behavior of the individual, these assumptions will not ultimately be effective either as motivating devices or as goals in the design of culture. . . . We may console ourselves with the reflection that science is, after all, a cumulative progress in knowledge which is due to man alone, and that the highest human dignity may be to accept the facts of human behavior regardless of their momentary implications" (11, p. 449).

Implicit in this quotation is the view that this approach involves no value judgments by the scientists who conduct experiments in controlling cultural design and modifying human behavior. In fact, Skinner elsewhere states explicitly that "our problem is not to determine the value or goals which operate in the behavior of the cultural designer; it is rather to examine the conditions under which design occurs" (11, p. 433). However, it does not seem clear, at least to me, that Skinner has adhered to this position. In spite of his assertion to the contrary, the choice of survival as a criterion for evaluating control, and the choice of a science of human behavior as mediating mechanism in deciding with respect to alternative courses of action, appear very clearly to be value judgments. Furthermore, with the literary license allowed to the novelist, Skinner in *Walden Two* has exercised wide latitude in this respect and thereby has revealed the dangers that arise when, in a life situation, the psychologist does, in fact, implement the view that his science makes him the architect preeminent of the utopian way of life.

There occurs, for example, a discussion of the community educational program. A visitor, named Castle, raises a question concerning student motivation. "Why," he asks, "do your children learn anything at all? What are your substitutes for our standard motives?"

To make clear the issue under consideration requires, unfortunately, a somewhat lengthy quotation from Skinner's novel, which goes on as follows (10, pp. 101-102).

" 'Your "standard motives"—exactly,' said Frazier. 'And there's the rub. An educational institution spends most of its time, not in presenting facts or imparting techniques of learning, but in trying to make its students learn. It has to create spurious needs. Have you ever stopped to analyze them? What are the "standard motives," Mr. Castle?'

" 'I must admit they're not very attractive,' said Castle. 'I suppose they consist of fear of one's family in the event of low grades or expulsion, the award of grades and honors, the snob value of a cap and gown, the cash value of a diploma.'

" 'Very good, Mr. Castle,' said Frazier. 'And now to answer your question—our substitute is simply the absence of these devices. We have had to uncover the worth-while and truly productive motives. . . .'

" 'We made a survey of the motives of the unhampered child and found more than we could use. Our engineering job was to preserve them by fortifying the child against discouragement.'. . ."

Following a description of the use of "conditioning" in building up tolerance to discouragement, the founder of *Walden Two* goes on to say, " 'Building a tolerance for discouraging events proved to be all we needed. . . The motives in education, Mr. Castle, are the motives in all human behavior. Education should be only life itself. We don't need to create motives. We avoid the spurious academic needs you've just listed so frankly, and also the escape from threat so widely used in our civil institutions. . . We don't need to motivate anyone by creating spurious needs.' "

Skinner uses here, of course, a device commonly employed by both literary men and expert propagandists in lulling the reader into at least the provisional acceptance of his viewpoint. It is that of molding attitudes by the choice of appropriate adjectives, illustrated in the quotation by the phrases "the snob value of a cap and gown," "the cash value of a diploma," and most of all by the repeated reference to "spurious needs."

Social Science and Social Reform

The last of these phrases, "spurious needs," brings into relief the situation that has produced both the title of this address and its content. This, briefly, is the increasing tendency on the part of the psychologist to inject value judgments in a manner that makes it increasingly difficult, especially for the layman, to determine when the psychologist is dealing with facts and principles derived from experiments, or when he is merely presenting his own value judgments (13). It has, in other words, become increasingly difficult to know when the psychologist speaks with the authority of science, or when he is playing the role of the social reformer while clothed—or even disguised—in the garb of the scientist.

In saying this, I am, naturally, not denying the right of the psychologist to his opinion—to his own value judgments. He, as every other free man, is entitled to believe that a cap and gown is, indeed, a stigma of snobbery; that a diploma is prized only for its cash value; that money is crass; that, as Rogers believes, religion, and also Freud, are to be criticized for permeating our culture with the false concept that man is sinful (12); that prejudice and discrimination are used by dominant groups to defend their vested interests (14), and so forth. As a citizen, the individual psychologist is free to express any such opinion, regardless of how unpopular it may be among his professional colleagues or among the mass of people in the culture of which he is a part. It is not his privilege, however, to clothe the source and personal nature of such opinions in the language or form of scholarly writing to the point where it would appear that they are the outcome of scientific inquiries.

Reference to *Walden Two* as a device for presenting this issue does not reflect the opinion that Skinner has been particularly remiss in this respect, in comparison with other psychologists. This fictional representation of his personal views by a notable and conscientious scientist merely provides a springboard for the discussion of a major issue in psychology. It is an issue that grows in significance with the multiplication of publications where the failure to distinguish between conclusions supported by experimental evidence and those representing personal value judgments becomes a medium for the support of cultural practices or changes deemed to be desirable by the scientist.

The frequency with which this occurs lends support to the opinion that many psychologists have reverted to Plato's conception of method, as stated in *Phaedo*, namely, "This was the method I adopted: I first assumed some principle, which I judged to be the strongest, and then I affirmed as true whatever seemed to agree with this . . . and that which disagreed I regarded as untrue." The fact that, in most instances, the individual psychologist is not engaged in the patterning of an entire utopia, but rather in what Popper in *The Open Society and Its Enemies* (15) has called "piecemeal social engineering," does not diminish the seriousness of the situation under discussion, especially in an era that has raised the psychological expert to a level of considerable influence.

Essays in Piecemeal Social Engineering

Many examples of this situation can be cited. A thought-provoking article by Gardner Murphy, entitled *Human Potentialities*, furnishes one such illustration. Here, Murphy formulates five basic principles for "permitting the discovery of human potentialities," including among these, as a negative principle, *to avoid the competitive*. "Not," he wrote, "because competition is always bad, but because it frustrates and benumbs those who fail, and because for those who succeed it can at best give only the ever iterated satisfaction of winning again and again. In this direction lies, of course, a convenient way of maintaining a status minded society; but I am speaking of something quite different, namely, the release of human potentialities" (16).

Accepting Murphy's statement that he is interested primarily in the release of human potentialities, there still arises the question whether there are, indeed, facts available to support the use of the word *principle* instead of *judgment* or *opinion* in the context of his statement. Furthermore, the reference to "status minded" society introduces at least an implication that "competition" is a socially undesirable practice, as well as a handicap to the full and healthy development of the individual.

Examination of the literature—particularly that of social psychology—indicates that competition is quite frequently treated as though it has been demonstrated with considerable certainty that this is a noxious cultural practice. In addition, by associating capitalism with competition, onus is reflected on the capitalistic system, as compared with other and, by implication at least, superior economic and social systems. Thus, according to Newcomb, the higher frequency of exposure to failure, threat, and insecurity that exists where importance is attached to competitive success makes it "no wonder that psychiatrists like Alfred Adler found feelings of discouragement and inferiority prominent in the neuroses of Western society" (17). In a somewhat broader context, Asch states the requirements that distinguish between a "society of atoms, each arrayed against all, organized on the predatory principle of *homo homini lupus* and one organized around the idea of a community of men." The former, it is made clear, is one built on the "calculation of private profit." Only an inferior brand of social organization can be anticipated from an "ego-centered thesis" that "describes the balance achieved in society as an uneasy and antagonistic mutual limitation of each by all" and that "reduces every trace of solidarity to the pattern of relations in the business market" (18).

How many facts, from how many studies, are available to support such judgments with respect to the individual and social roles of competition? Newcomb's reference to Adler's statement concerning the frequency of neuroses in Western (competitive and capitalistic) society merely raises again the perennial questions concerning what constitutes neurosis"; concerning the amount and quality of research underlying psychiatrists' dicta, and even concerning the nature of the sample observed by the psychiatrist. The last of these questions is neatly disposed of in the reply given to the query "Whom has the psychiatrist been observing?" in a humorous but nevertheless challenging book entitled *How to Lie with Statistics*. "It turns out," it is pointed out, "that he has reached this edifying conclusion from studying his patients, who are a long, long way from being a sample of the population. If a man were normal, our psychiatrist would never meet him" (19).

Perhaps the situation with respect to research on competition versus cooperation is not quite as bad as this. However, the fact remains that studies bearing on the effects of competition on the individual and on groups are few in number. Furthermore, the size and nature of the samples involved in such studies, the restricted and frequently artificial settings in which they are conducted, the manipulation of theoretical concepts and experimental variables, and so forth, make it quite impossible to derive broad value judgments pertaining to the role of competition in social progress. Available experimental findings do not provide grounds for discarding lightly the opinion, expressed in a prophetic dissent by Justice Holmes of the Supreme Court of the United States, that competition (between groups as well as between individuals) is a social advantage since it "is worth more to society than it costs" (20). Certainly, the hypothesis that competition—reaching even the dimensions of conflict—contributes to individual and group progress cannot be abandoned. This, in fact, is the position taken with respect, at least, to the social role of conflict in industry by a number of contributors to a recent book, *Industrial Conflict*, edited by Kornhauser *et al.* (21).

This reference to industry brings to mind another illustration of the presentation of value judgments unsupported by facts derived from research. There has been considerable thought given to the role of the union, in comparison with that of other social organizations, in providing "substitute" satisfactions for wants and needs that are presumably frustrated by the job conditions under which people work in modern industry. Writing within the context of a scholarly work, Krech and Crutchfield state with conviction that "*the labor union, by and large, can better meet most of the workers' needs and demands than can other organizations.* As we have seen . . . most social organizations will generally reflect the major needs of its members, and labor unions will therefore be more 'tailored' to the needs of the workers than will religious organizations or other less homogeneously composed social organizations" (22, italics mine).

In 1948, at the time this statement appeared, there was little available in the way of research findings bearing on the workers' perception of other social organizations (apart from the industrial plant) in comparison with their perception of the union. So far as religious organizations are concerned, there were not, to my knowledge, any facts that would support or disprove the conclusion reached by Krech and Crutchfield.

Studies conducted since 1948 do not show that workers themselves perceive the union as the prime medium for satisfying most of their needs. Thus, in a study of a teamsters union, by Rose, 75 percent of members referred to "getting higher wages," and 31 percent to getting "job security," as a purpose of the union (23). *No other single purpose is mentioned by as many as 20 percent of the workers involved.* Similar findings, in other studies dealing with the worker's perception of the union (24), likewise throw serious doubt on the view that the union does or can satisfy the needs for participation, for self-expression, for self-respect, for status, or a host of other psychological and social needs better or more fully than do other types of social organizations.

There is still little, if any, evidence bearing specifically on the question whether labor unions can or will be more "tailored" to the needs of workers than will religious organizations. It seems true, as Krech and Crutchfield contend, that unions are, in fact, assuming accessory functions of the type that enlarge the potential for the satisfaction of more and more needs of its members. As is also pointed out by Krech and Crutchfield, this is likewise true of religious organizations. They provide no evidence that one is doing this to a greater extent or with better results than the other. Furthermore, although current research on dual loyalties—for example, to the union and the religious organization—points to the fact that each organization may better satisfy some specified need, findings do not in any sense settle the question whether either is or can be better "tailored" to provide direct or "substitute" satisfactions for most needs.

In using this illustration, I am not, for the moment, concerned with the evaluation of the role of either the union or of religious organizations in the life of the individual and in modern society. I am concerned with treatment of the roles of these and of other social organizations by psychologists in a manner that confuses theory or value judgments with facts—in a manner that may, with or without intent, mold the attitudes of the reader or student with respect to social institutions rather than enlighten him with respect to their roles as revealed by research. The finding, reported in a recent study by Keehn, that the resemblance within a group of well-known psychologists ($N = 27$) was confined to high homogeneity with respect to a continuum of "humanitarianism and anti-religionism" (25) perhaps lends special pertinence to the illustration under consideration.

Many illustrations of premature and also biased generalizations from relatively little in the way of facts are to be found in industrial applications of psychology that, as may be suspected, are of special interest to me. Thus, earlier discussions of the effects of repetitive work, and also current discussions of automation have suffered both from an absence of historical perspective and from the "naturalistic fallacy" in which subjectively determined goals and moral values are confused with the empirical methodology and outcomes of scientific research (26).

A necessarily brief illustration from another area of research and application may help to reveal the wide scope of the problems under discussion in this article. In a volume entitled *Motivation and Personality*, Maslow takes the position that "science is based on human values and is itself a value system" (27, p. 6). Acting on this premise, he has described a utopia, called *Eupsychia*, characterized by the fact that all men are psychologically healthy. Essentially, according to Maslow, this means that "the inhabitants of Eupsychia would tend to be permissive, wish-respecting, and gratifying (whenever possible), would frustrate only under certain conditions . . . and would permit people to make free choices wherever possible. Under such conditions," adds Maslow, "the deepest layers of human nature could show themselves with great ease" (27, p. 350).

Here Maslow appears to accept what Skinner has described as a dominant view characterizing the theory and practice of psychotherapy (expressed earlier in the primitivism of Rousseau), namely, that man is essentially good and kind and is corrupted only by social forces imposed from without. Thus, Rogers, the high priest of psychotherapy, takes issue with Freud's view (28) that man's basic nature—the *id*—"is primarily made up of instincts which would, if permitted expression, result in incest, murder, and other crimes" (12, p. 56). The contrary, Rogers contends, is the fact! "One of the most revolutionary concepts to grow out of our clinical experience," he writes, "is the growing recognition that the innermost core of man's nature, the deepest layers of his personality, the base of his 'animal nature,' is positive in character—is basically socialized, forward-moving, rational and realistic" (12, p. 56). The goal of psychotherapy therefore naturally becomes that of providing a client-centered, permissive atmosphere that leads to *adjustment* through the revelation—by the individual to himself—of the essentially "self-preserving and social inner core" of his personality (29).

Which of these views—that of Freud, or that of Rogers—can we accept as scientific truth? In what measure are the tremendous structures of psychoanalysis and psychotherapy built on a foundation of empirically established facts? And to what extent can we accept adjustment itself as a prescription for living "as a socially desirable goal?" Or is their justification for Lindner's view that the whole concept of adjustment "is a mendacious lie, biologically false, philosophically untenable, and psychologically harmful"—which, according to Lindner "disregards many if not all the pertinent facts of human nature" and represents "an untruth that is rendering man impotent at a time when he needs the fullest mastery over his creative abilities" (30).

The Scientist and His Moral Values

Whether this is true or not (31), the sad fact is that the immense superstructure of psychological practice often rests on a foundation of scattered, splintered, and tinderlike data that could fall apart with the most meager essays in the way of further exploration through the use of available scientific techniques. Psychologists and psychiatrists alike seem loath to acknowledge this. Only too often we seem possessed—not by an appropriate and deep sense of humility—but, instead, with an urge to substitute our value judgments—frequently uncontaminated by facts—for those held by others and as perhaps expressed by colleagues in related fields of economics, history, political science, philosophy, religion, and so forth. Like Scaphio and Phantis in the delightful comedy *Utopia Ltd.* by W. S. Gilbert, we seek to enter the world of affairs to the voice of a chorus that sings (32)

> "O make way for the Wise Men!
>   They are prizemen—
>
> . . . .
>
> They're the pride of Utopia—
>   Cornucopia
> Is each in his mental fertility.
> O they never make a blunder,
>   And no wonder,
> For they're triumphs of infallibility."

It is possible that in this paper—and also in my earlier publications—I may appear to have clothed myself in the mantle of the "wise man." It is unquestionably evident that much if not all that I have said here is in the nature of value judgments. In fact, I make no claim to the scientific authenticity of my judgments. Furthermore, this article does not purport to set up a scientific system of moral values, or even to support the position that this can be done.

Nevertheless, moral values are involved, and these require serious thought whenever psychologists turn their attention to newer developments in the way both of the theory and applications of the science of human behavior. This seems the occasion to recall the description, by Pliny, of the activities of the clothiers of Rome who met in the Forum in the autumn of each year and whose activities made *caveat emptor*—let the buyer beware—the expression of bitter experience on the part of the Romans (33). The very fact that the infant science of human behavior can already make important and useful contributions to human welfare does not entitle us to play the role of the architects preeminent of the new utopia.

We are not privileged to let our individual moral values—instead of hard facts—set our standards of conduct as scientists. We cannot conscientiously permit even a despair of finding ethical absolutes to lead us, in the words of Keckskemeti, to "smuggle them in behind intellectual, psychiatric, and political screens" (34). There is no time better than now to recall the forceful appeal by A. V. Hill that "scientists should be implored to remember that, however accurate their scientific facts, their moral judgments may conceivably be wrong" (35). Let us take pride and courage in the dedication of our work as scientists to the cause of mankind—to defending and enhancing the worth of the human being (34, p. 371). We must, nevertheless, simultaneously keep constantly in mind the necessity for clearly separating our thinking and wishes with respect to ordinary affairs from the "critical habits of thinking" (35) that characterize the true scientist and establish the inherent integrity of a science.

REFERENCES AND NOTES

1. *Practical Standard Dictionary* (Funk and Wagnalls, New York, 1924).
2. S. Butler, *Erewhon* (1872; reprinted by Dutton, New York, 1917), and *Erewhon Revisited* (Richards, London, 1901).
3. E. Bellamy, *Looking Backward* (Houghton Mifflin, Boston, 1898).
4. C. de Bergerac, *L'Autre Monde ou les Etats et Empires de la Lune et du Soleil* (about 1655; reprinted by Le Cercle du Livre de France, New York, 1953).
5. M. Arnold, *Culture and Anarchy* (1869; reprinted by Macmillan, New York, 1925), Chap. 1.
6. *Rubáiyát of Omar Khayyám* (translation by E. Fitzgerald, ed. 1, 1859; reprinted by Random House, New York, 1947), verse 73.
7. J. J. Rousseau, *Discours sur les Sciences et les Arts* (1750; reprinted by Oxford Univ. Press, London, 1946); *Discours sur l'Origine et les Fondements de l'Inégalité parmi les Hommes* (1755; reprinted by Oxford Univ. Press, New York, 1922); *Emile* (1762; reprinted by Didot, Paris, 1851).
8. F. R. de Chateaubriand, *René* (1800); *Atala* (1801); *Les Natchez* (about 1800), published in *Oeuvres Completes de Chateaubriand* (Garnier, Paris, about 1850).
9. A. Huxley, *Brave New World* (Doubleday Doran, New York, 1932).
10. B. F. Skinner, *Walden Two* (Macmillan, New York, 1948).
11. ———, *Science and Human Behavior* (Macmillan, New York, 1943).
12. See particularly C. R. Rogers, "Some directions and end points in theory," in O. H. Mowrer, *Psychotherapy: Theory and Research* (Ronald Press, New York, 1953).

13. It is apparent that I here (as also elsewhere in this paper) distinguish between *fact* and *value* and, at least by inference, reject the view, appearing in current discussions of theory of knowledge, that facts are in themselves value judgments. Actually, I do *not* accept the view that *existential* and *normative* propositions are equivalent—that *scientific* and *ethical* statements are basically similar [G. Lundberg, "Semantics and the value problem," *Social Forces* 27, 114 (1948)]. By contrast, I am inclined to accept the view, as expressed by C. Kluckhohn, that although existence and value are intimately related and interdependent, they are—"at least at the analytical level—conceptually distinct." However, a detailed discussion of this controversy is not appropriate in this paper. The reader interested in a detailed discussion of theoretical considerations in this area is referred to publications cited here, particularly reference 26, and, in addition, to a chapter on "Values and value orientations in the theory of action: an exploration in definition and classification," by C. Kluckhohn *et al.*, in *Toward a General Theory of Action*, T. Parsons and E. A. Shils, Eds. (Harvard Univ. Press, Cambridge, Mass., 1951), pp. 399-433.
14. See particularly G. W. Allport, *The Nature of Prejudice* (Addison-Wesley, Cambridge, Mass., 1954) and G. Saenger, *Social Psychology of Prejudice* (Harper, New York, 1953).
15. K. R. Popper, *The Open Society and Its Enemies* (Princeton Univ. Press, Princeton, N.J., 1950).
16. G. Murphy, "Human Potentialities," *J. Soc. Issues Suppl. Ser. No. 7* (1953), pp. 14-15.
17. T. M. Newcomb, *Social Psychology* (Dryden, New York, 1950), p. 27.
18. S. E. Asch, *Social Psychology* (Prentice-Hall, New York, 1952), p. 316.
19. D. Huff, *How to Lie with Statistics* (Norton, New York, 1954), p. 19.
20. B. Aaron, "Changing legal concepts in industrial conflict," in A. Kornhauser, R. Dubin, A. M. Ross, *Industrial Conflict* (McGraw-Hill, New York, 1954).
21. A. Kornhauser, R. Dubin, A. M. Ross, *Industrial Conflict* (McGraw-Hill, New York, 1954).
22. D. Krech and R. S. Crutchfield, *Theory and Problems of Social Psychology* (McGraw-Hill, New York, 1948), p. 548.
23. A. M. Rose, *Union Solidarity* (Univ. of Minnesota Press, Minneapolis, 1952).
24. See M. S. Viteles, *Motivation and Morale in Industry* (Norton, New York, 1953), Chap. 18.
25. J. D. Keehn, "The expressed social attitudes of leading psychologists," *Am. Psychol.* 10, 208 (1955).
26. See particularly A. J. Ayre, *Language, Truth and Logic* (New York Univ. Press, New York, 1936); G. E. Moore, *Arguments against Ethical Naturalism* (Northwestern Univ. Press, Evanston, Ill., 1942); and P. Keckskemeti, *Meaning, Communication, and Value* (Univ. of Chicago Press, Chicago, 1952).
27. A. H. Maslow, *Motivation and Personality* (Harper, New York, 1954).
28. See particularly S. Freud, *Civilization and Its Discontents* (Hogarth, New York, 1953).
29. Studies of "feral" children and other investigations bearing on the "essential nature of man" are summarized in a recent volume by C. Leuba, *The Natural Man* (Doubleday Doran Papers in Psychology, 1954).
30. R. Lindner, *Prescription for Rebellion* (Rinehart, New York, 1952), p. 12.
31. Lindner might find support for his views on the nonresistant adjusted man in a recent comparison, by the biologist H. W. Stunkard, of the sources of degeneracy among parasitic animals and inhabitants of the societies of ants and bees with the loss of freedom in the cult-controlled welfare state of mankind ["Freedom, bondage, and the welfare state," *Science* 121, 811 (1955)].
32. W. S. Gilbert, *Plays and Poems* (Random House, New York, 1935), p. 588.
33. M. Beard, *A History of the Business Man* (Macmillan, New York, 1938), pp. 40-41.
34. P. Keckskemeti, "The psychological theory of prejudice," *Commentary* 18, No. 4, 359 (1954).
35. A. V. Hill, "The social responsibility of scientists," *Bull. Atomic Scientists* 7, 371 (1951).

PANEL DISCUSSION

Clinical vs.
Actuarial Prediction

Clinical And Actuarial Prediction in a Setting of Action Research

NEVITT SANFORD

When I first looked over Paul Meehl's moving account (2) of his conflicts about clinical versus statistical prediction, I thought of course I would be able to say something helpful, something which if not therapeutic would at least be comforting. I myself had not been troubled by this particular conflict, supposing as I did that statistical prediction was merely a tool for demonstrating, or testing the generality of, what one knew already. Now, since I have looked into this matter somewhat more carefully, I must admit that Paul Meehl has me worried. If there are not places for both clinical and statistical prediction, if the two cannot be reconciled, then I have to look forward to the imminent splitting of my personality. Hopefully, then, but not without anxiety, I am trying to figure out how in my scheme of things clinical and statistical methods are related, or kept separate, integrated or confused.

I would suggest at the start that what divides the clinicians and the statisticians—what they get passionate about at any rate—are not so much differences about the best way to perform a given task, as differences in more general outlook—perhaps even in temperament. The arguments are very likely to concern what ought to be predicted, what predictors to use, what level of predictability is possible or desirable, what is going to be done about the predictions once they have been made.

What I propose to do now is consider some of our activities at Vassar, with attention to such issues as these, and in the light of some of the arguments that have been advanced in favor of the two kinds of prediction.

We are trying to predict withdrawal from College, by methods that are strictly actuarial—even "blindly empirical." How can we justify this in the eyes of our clinical friends?

In our circumstances this is a far less expensive proceeding than any clinical one we might use to accomplish the same thing. A battery of some 1100 items having been given to 4 entering freshmen classes, the matter of finding predictors of withdrawal is a straight-forward machine operation. We do not want to take the time to study admissions data, look up students who have dropped out, or to interview a sample of entering freshmen and on that basis guess who will withdraw.

Moreover, the evidence is that our mechanically constructed device will score more hits than would the usual "clinical" procedure, e.g., guessing on the basis of an interview, or perhaps a few records and tests. (It will have to go some, however, to do better than the Admissions Committee, who predicted that no entering students would withdraw and, as far as the freshman year was concerned, was correct in about 90% of the cases. The Committee proceeded clinically, I suppose.)

Consider the difficulties of making clinical predictions of a criterion such as we are considering. Say that I know the subjects well, chiefly on the basis of intensive interviews. I would be biased—probably in the direction of leniency—and I would be confused. I would think of so many hypotheses favoring one or the other action, withdrawing or remaining in college, that I would feel quite lost when it came [illegible] the matter of assigning weights. (Harold Webster tells me that a clinician who can think of many factors which seem to have some association with the criterion would probably do best if he just gave them all the same crude weight. I'm sure that we often over-weight an interesting psychodynamic factor, or else over-compensate for a tendency to do so by supposing that intelligence is of paramount importance.)

If I knew the subjects well, I would probably be thinking about the relative strengths of variables in a given individual, and about how the variables related one to another, rather than about group norms for any of these variables. The chances are that I would know little about the situation in which the criterion behavior occurs. (Is it not often the case, in clinical prediction studies, that they involve either clinicians who have only vague notions about the criterion or else people such as deans or admissions officers who know the criterion but are not very good clinicians?) Actually, in the case of an entering freshman whom I had learned to know well, I might have very good notions about whether or not I wanted to take him along on a cruise, when I would be called upon to anticipate his behavior in a thousand situations, and still be quite unsettled about such a dichotomous criterion as dropping out versus not dropping out of college.

So it seems to me that in the present instance it ought to be granted that statistical methods can do better.

It should also be recognized that such a predictive device as we are working on can have little practical value when it is taken by itself. Its applicability is both limited and dubious. We already know that what holds for Vassar does not hold for certain other colleges, and that what holds for freshmen does not hold for sophomores. And it is recognized that if there should be changes in the way the college manages its students the whole thing would have to be done over again.

It must be recognized, too, that however good our statistical predictions might be, the college will not adopt a statistical policy with respect to admissions.

Much of the passionate rejection of empirically derived tests, by clinicians and humanists, is based on a fear, sometimes justified, that what has been derived actuarially will be applied collectively. It does seem that strong adherents of actuarial methods tend to be institution-centered, while clinicians tend to be individual-centered.

If our predictive device has little practical value, it would seem to have even less scientific value. I am assuming that we merely pull items, and cross validate in successive groups. We establish a close relationship between, let us say, "tendency to drop out of Vassar in the freshman year," as measured by tests, and dropping out of Vassar in the freshman year. We define no psychological variables, state no hypotheses, invoke no theory. This kind of thing is actually done. Since actuarial prediction of socially defined criteria became the order of the day, the study of personality for its own sake has been rather neglected. I know of one research organization that was founded with the object of studying personality but which became converted to actuarial prediction of practically important criteria, and where discussion of psychology is no longer heard. Only methodology is discussed.

Why then do we bother with this apparently trivial exercise? Can we yet manage to derive some scientific and practical value for it?

For one thing, we will make it serve an exploratory function. A study of the scales and items which separate drop-outs from remainers will give us some notions of what is going on. It will yield suggestions about the college as well as about processes in the students. Once again, it is a much less expensive procedure than making a sort of anthropological investigation of the college and a clinical investigation of leavers.

These investigations might, however, turn up personal and situational factors that did not over-lap entirely with those suggested by our "blindly empirical" approach.

At any rate, with such factors in mind we will be in a position to formulate and to test some hypotheses. If students with tested characteristics a, b, c, and d are dropping out of College A, because of conditions x and y in that college, then we will predict that students of this sort will not drop out of College B, where these conditions do not obtain—nor out of College A, should these conditions be changed.

As a matter of fact, to speak of practical matters, discussion with the college of results—suitably interpreted—of the statistical prediction study might conceivably set in motion a process of self-criticism which would lead to changes in conditions x and y.

The early identification of probable drop-outs might be the basis for starting a counseling program that would reduce withdrawals among those students, if any, whose interests were better served by remaining in the college. If students dropped out despite this counseling, one should have a pretty good understanding of why they did.

A study of "false positives" would be particularly interesting, for the light it might shed on education at the college. One always hopes, of course—and with little doubt that the hope will be realized—that the predictions will not be *too* good.

What would a true clinical approach, in our situation, entail? At the least, it would seem, 6 or 8 hours of interviewing and testing, with a sample of entering freshmen, in order to arrive at a "dynamic formulation" of each case. Students would undoubtedly be changed by this clinical work, quite possibly in a way that would make them less likely to drop out of college. Prediction would have to take this circumstance into account. This whole business could, quite conceivably, be put into an actuarial table or equation; but this would be useful, of course, only in situations where this same program was in effect.

By the time one had completed this clinical work, he would very probably have lost interest in whether his subjects dropped out of this college or not. Other, broader, aspects of their future lives would, by and large, be seen as much more important. I assume that no clinician would undertake such a project *just* to see whether he could predict as well as the statisticians.

At the conclusion of this work the clinician might understand the student well enough so that he could explain some of his processes to him—if this seemed to be in the student's interest. (And this, by the way, is a pretty good test of one's understanding of another person. It is the kind of knowledge which when verified and generalized is of the very essence of the psychology of personality.)

And the clinician might feel an ethical obligation to pass some of his knowledge along to the student. No one has a better right to it. This may well play hob with the prediction concerning withdrawal. But, on the other hand the clinician would have the advantage of what might well be the best predictor of all, that is, what the student says—to a trusted counselor—he is going to do.

Such a clinical approach might be of very considerable practical value, assuming that the concern is with education and welfare and not merely with drop-outs. The college could learn something about itself; and the student's decision to withdraw or not might be of a more considered kind.

Consider in this connection the Tavistock Institute's work with Company X. Three clinicians and three officials of the company observed, in a group discussion situation, applicants for positions; then they divided the interviewing so that each applicant was seen by two clinicians and two officials including the head of department in which a given applicant would work, and it was later decided in conference who was to be selected. A poor way, it would seem, to determine what kind of man made good at Company X, but an excellent way, as it turned out, to reduce turnover of highly trained personnel and to improve morale in the company. The officials were learning quite a lot about themselves and quite a lot of psychology.

If the clinician were to put in 6 or 8 hours per subject, in the above project, it would be surprising if he did not end up in a counseling relationship with some of them, making referrals in other cases.

As a counselor or psychotherapist he would assume, with Kluckhohn and Murray (1) that each client or patient was "in certain respects a) like all other men, b) like some other men, c) like no other man." He would lean most heavily upon general laws of human functioning, however crudely these were formulated, and next most heavily upon his conceptions of syndromes, patterns or types that were more or less common. He would try to order to these general laws and conceptions the generalizations he would have to make about the unique productions of his client.

He would expect to receive relatively little help from tests. Of course, there are no objective tests for more than a small fraction of the variables with which he would have to deal. He would find other people's formulations on the basis of projective tests interesting, but not a very practical investment of time and energy; he would have to make his own formulations in any case, and he would probably consider that he was in a better position to do this than was the projective tester. He would regard empirically derived predictors of success in psychotherapy as a useful check upon his own judgment. He could not take them too seriously, considering as he would that success in psychotherapy is still undefined and that the predictors probably did not apply to *his* psychotherapy anyway. He would have no great difficulty—assuming good training—in recognizing that he had a tough case on his hands, but numerous other considerations might outweigh the dubious prognosis in his decision to try and help the person.

In sum, I seem to have come out in favor of interaction between clinical and statistical methods, on the grounds that each can be supportive of the other. They need be competitive but rarely, since for a given task one or the other can usually be judged to be better suited.

In scientific work, the major role of statistical prediction is still to demonstrate what has been observed in clinical or experimental situations. It has been suggested here, however, that the development of empirical predictors of socially defined criteria may also have an important exploratory function; it may suggest hypotheses of general scientific interest. The fact that so much effort is directed to the statistical prediction of criteria which are socially important but scientifically dubious is hardly a criticism of the method itself. It is up to those who are primarily interested in personality functioning to define and to estimate the variables which for them are of fundamental importance, so that the great potency of statistical prediction may be directed to these.

When it comes to practical work the thing to emphasize is the huge gap between what is known or can be known from actuarial prediction and what needs to be known and considered in order to determine wise policy. In clinical work with individuals the matter is quite clear. The clinician should be grateful for whatever objective test results that can, without too much expense, be placed in his hands, but he can do no more than consider these in their place among a host of other things which he must judge and act upon. Matters are not very different from this in the case of applications within an institutional setting. One might hope that here too the psychologist will take part in the analyzing and the judging of the whole complex of affairs within which his predictive devices have a relevant place.

REFERENCES

1. KLUCKHOHN, C. AND MURRAY, H. A. *Personality in Nature, Society, and Culture.* New York: Knopf, 1948.

2. MEEHL, P. E. *Clinical vs. Statistical Prediction.* Minneapolis: Univ. of Minn. Press, 1954.

Clinical Versus Actuarial Prediction

CHARLES C. McARTHUR

Our question is, "Which predicts better, clinical or actuarial methods?" The correct answer is, "We don't know; no one has done the experiment." The moral is, "Somebody ought to!"

I know there have been experiments purporting to answer this question. They just seem for the most part so poorly designed that they are irrelevant.

How should a relevant experiment be designed? Well, the general rule is that both clinician and actuary should be given every opportunity to show their wares. That's the only possible way to match them. If, with apparent scientific sophistication, you hold the conditions under which the actuary and the clinician must predict "constant," one or both men will be handicapped by the conditions you prescribe. How can you say "how much" of a handicap each man suffers, or how to make his opponent's handicap "equal"?

Years ago, we had a similar problem in intelligence testing. In the early days of the testing movement, it was thought that the way to be "fair" to all the children tested was to repeat precisely the same external ritual, with the examiner working in the same office, with the same lighting, introducing himself and the test in the same way and giving instructions verbatim. The I.Q.'s so derived were presumed to be directly comparable. Alas, they were not. What was soon learned about intelligence testing was that true comparability could be had only when the examiner varied his behavior appropriately from child to child, so as to obtain for each child the maximally favorable conditions. When each child was "given all the breaks," the resulting I.Q.'s could be justly compared.

So it is with our question. If we really want to know how the clinician and the actuary compare, we have to let each man

(a) use the data of his choice,
(b) make the analysis of his choice, and
(c) make the predictions of his choice.

Now, I'm not an actuary and I'm speaking to actuaries, so I'd better let them make their own choice of data, analyses and predictions. It is about the clinical half of this contest that I feel entitled to speak, if only because, at the Study of Adult Development, we have recently gathered some experimental observations on the way people gather good and bad clinical predictions. I would like therefore to review the proper choices of data, analysis and predictions for the clinical half of a good clinical versus actuarial experiment.

* * *

If we want to make good clinical predictions, what kinds of data† will be the data of our choice?

We want plentiful data. Plentiful enough to make us feel that we may have an adequate sample of all kinds of our subject's behavior. Those of us who earn a living as working clinicians go day after day, year after year, jumping to premature conclusions on inadequate evidence. That's what we're paid to do. All the same, when we do experiments for the advancement of knowledge, we are forced to accept the stern reminder of Robert White that "An attempt to cut the testing schedule below ten to fifteen hours with each subject is merely a proposal to sabotage the research." A more usual battery for experimental purposes would run to thirty or forty hours.

We want various data. It is almost indispensable to watch one subject interacting with at least half a dozen examiners. It *is* indispensable to sample his behavior at all psychic "levels." Projective devices are a must but so are observations of S. performing workaday acts in his everyday setting.

We want overlapping data. We'd best see our man tackling comparable problems under very different conditions, with different examiners, different degrees of stress, in different contexts.

We want open-ended data. The ratio of the subject's talk to the examiner's talk should be at least ten to one.

We want fully recorded data. That is another lesson from intelligence testing. More and more, the by-products of a test situation turn out to be more useful than the measurement that was the historical purpose of the test. A Wechsler-Bellevue recorded verbatim, the irrelevant remarks being recorded most scrupulously of all, tells us many times as much as the I.Q. or even the sub-test profile of scores. White has made this point well, at the same time explaining why we insist on obtaining overlapping data. "Our problem-solving test," he points out, "will perhaps also be a test of frustration tolerance, a test of control over anxiety, a test of level of aspiration, or a situation that happens to mobilize an infant traumata, *and our report on its results must include as much of this information as can be observed.*" The rhetorical italics are mine.

---

† Quite typical of the problems of communication across the two frames of reference, clinical and actuarial, is the fact that this talk was prepared with no awareness that the word "data" contained any ambiguity. At luncheon before the panel discussion, the writer became aware that when he says "data" he means "contents of verbal or behavioral acts" and that when actuaries say "data" they mean "scores." Neither usage is perfectly exclusive but the difference is gross enough to tangle communication badly.

* * *

If we have now collected the right kind of data, we may be in a position to take our second step and ask what should be the clinical analysis of our choice. And there *is* one best analysis. I, too, have heard the rumor that each clinician uses the method that is his personal favorite. The rumor may even be true; tastes vary, though science be constant. Nonetheless, both logic and empirical validation identify *one* best technique of clinical analysis. This technique is neither intuitive, as rumor so often has it, nor a Mystery, nor is it unavailable to actuaries. You see, the clinical analysis of choice is nothing but the application of the Scientific Method!

I am not the only clinician who has this idea. "The diagnosis of each personality is," according to White, "a miniature scientific experiment." Meehl would also, although with some caution, accept the idea that the good clinician makes " 'little special theories' the applicability of which is to one person."

Nor am I without evidence. At the Study of Adult Development, we recently asked a series of clinicians to formulate rich case data that was ten years old, and then to make postdictions of the subject's behavior during the ten years since the last recorded entry. We knew what the subject had been doing these last ten years; and, while our clinical prophet tried to guess what had happened, we all smugly sat around in a circle, "holding the book on him."

What all our clinical prophets did under these very trying validation conditions seemed to be to build from the data a clinical construct, a conceptual device, a "special theory applicable to one person," a *model* of that person, that made this statement on page 17 of the record consistent with that remarkable quotation back on page 14. Each datum became grist from which was ground a formulation of the premises governing *all* of S.'s behavior, the lifelong premises, the treasured self-consistencies with which the person being studied had learned to face the world. Each batch of data lent itself to hypotheses about the person, hypotheses that could be checked out against new data as the record progressed and could be revised with each successive cross-validation provided by turning another page. After conning all the data, the clinician possessed a fuzzy, but gradually sharpening, conceptualization of the man under study. "He seems to be the sort of a person who . . ." Then the clinician could make his predictions by doing imaginary experiments with the model. There would be paths down which the conceptualized person could effortlessly stroll, while there were alleys into which he simply could not be made to turn. And that was how good predictions got generated.

Tomkins has rigidly formulated this technique. Perhaps his most important statement is that we have to derive from the data itself *the very categories in which that data will be cast.* "In general," he warns us, "we do not know exactly what to look for. If we prejudge the categories of analysis, we may commit serious errors. What check, then," he asks, "have we on the adequacy of our selection of categories of analysis? It is our conviction that the logic of the individual's fantasy itself must be our ultimate criterion." End of quote. I would agree unreservedly. That facility at induction which enables him to derive for each new person studied a fresh set of categories that maximize the patterning of this particular set of data is the very hallmark of the good clinician.

Tomkins goes on, in a chapter that should be required reading for all graduate students, to specify how one can deduce from the data what categories were implicit in the mind of the subject himself. Nor are Tomkins' instructions vague or dependent upon intuition; he uses as his tool John Stuart Mill's canons of logic! Mill sets down rules like, "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause, or effect of the given phenomenon." Or else, "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common, save one, that one occurring only in the former, the circumstance in which alone the two instances differ is the effect, or cause, or an indispensable part of the cause of the phenomenon." And so forth, down through the Joint Method of Agreement and Difference, the Method of Concomitant Variation, Necessary Causes, Sufficient Causes, etc.

Furiously pedantic as they may sound, Mill's canons work. Suppose, for instance, your man tells one T.A.T. story in which the boy's mother wants him to practice the violin but the boy rebels, afterward feeling very guilty. Suppose in another story the mother wants the son to go to school but the boy quits school, again feeling guilty. Suppose a third story tells of another rebellious son who is now, however, being reconciled to his mother and doing as she wishes, and consequently this son becomes a great success and feels very happy. Mill and Tomkins would have us infer that, for this narrator, only a hero who does as his mother wishes may be permitted a happy ending. The category "obedient sons" will therefore play a dynamic part in our formulation of this man's personality and hence in what we can predict about his reactions to future events. For some other man whom we might study, this category could have absolutely no meaning.

I realize that clinicians don't usually think that systematically. The best empirical evidence about how clinicians actually think in practice is provided by Shneidman. After reviewing fifteen shockingly different systems for interpreting the same T.A.T. protocol, each system being offered by an "authority," Shneidman is able to discern a common set of steps in the clinical analyses. For most workers, the initial step is Charcot's: to look and look and look. They read and reread the data. The next stage seems to be "semi-organized notes" on repetitive or logically consistent patterns in the data. Then the criterion of internal consistency is applied and re-applied to trial hypotheses about the structure of the person's motives. Only in the end, when a diagnostic label is sought or if some one datum sticks out as incongruent with the rest, is any general psychological theory invoked. That was also true of our Study of Adult Development clinicians; theory came last. The one discussant we had who was embarrassingly inaccurate tried to deduce the behavior of the man he was discussing directly from the postulates of a general psychological theory. The successful prophets were those who remained inductive. None seemed to be as systematic as Tomkins would have them; but that only proves that the methods of analysis we use in everyday practice are less than ideal. It is probably true that we all could profit from more seminars entitled "The Diagnosis of Personality as an Hypothetico-Deductive Process."

* * *

Coming to the third portion of our clinical versus actuarial experiment, Tomkins' logic calls our attention to what sort of predictions clinicians should choose to make. If the categories in which we cast our data were those that logically arose from the data itself, then we have already decided what aspects of a particular case we can categorize, and hence we have unintentionally decided which aspects of the case we can predict. The clinical analysis has both this virtue and this liability: that it predicts what will be predictable . . . and what won't. An allied technique, usually called "thematic" analysis, has this same property. There are certain themes in which S. is very emotionally involved and it is these matters that we have most data on and can best predict. We have no basis for trying to predict any and all aspects of S.'s behavior.

Tomkins' formulation also gives us a second rule about our predictions: they must be contingent predictions. What we know is what Tomkins calls "the conditions for . . ." a certain behavior's appearing. "If S. perceives his boss as a nurturant elder, he will react by being ungrateful." If not, something else will happen. "If S. sees a woman as a sexual object, he will assume her to be evil, but if he perceives her as a supportive mother figure, he will assume her to be good." Always our predictions have the form "If . . . then . . ."

It follows that the usual experimental demand that the clinician predict multiple choice criteria, which look nicely objective but never state contingencies, almost certainly dooms the clinician to failure. It just isn't possible to say, in general, that "S. will be very aggressive." It *is* possible to say, "If S. sees the situation in this or that way, he'll be very aggressive." It is absurd to say, "S. will get well." It makes sense to say, "If his therapist can play this or that role toward S., S. will respond beautifully." Indeed, I wonder if it isn't more important to make these contingent predictions, not only because they turn out to be right more often, but also because they have more practical value.

Clinicians themselves don't seem to be aware of what predictions they can and can't make. Time after time, an excellent clinical analysis gets reported in terms of a rating scale, and so dooms itself to being invalidated on follow-up. When are we going to learn that we can't say "Mr. A. will be more aggressive than Mr. B." without specifying the conditions? Not only do our available methods of analysis prevent this, but it is quite likely that people just aren't made that way. Some of the most famous and recent and spectacular failures of really good clinical studies to stand up under cross-validation have arisen because of this one error.

I would insist, then, that any valid estimate of the accuracy of clinical prediction must permit the clinician to make contingent predictions and to limit himself to predictions about topics of his choice. I would hasten to add, however, that giving the clinician this liberty will not result in trivial, superficial, or safe and sure generalizations. The predictions the clinician can make relate to those very behaviors that have most importance of all, because they are the behaviors that matter to the subject himself.

* * *

So now we have reviewed three sets of conditions: the proper data, the proper analysis, and the proper predictions that must be had if we

101

are to learn whether the clinician be a prophet or a charlatan. Perhaps you see why I feel that none of the studies done till now are very relevant.

There are two sets of such studies.

Meehl has judiciously reviewed a set of studies that hold data and predictions constant while comparing two forms of analysis, one actuarial, one what the experimenters call "clinical." Perhaps the best known of these experiments is Sarbin's, though Meehl has located a dozen and a half more. The nature of the analysis is always insufficiently specified, but the piecemeal data supplied as a basis for prophecy always seems to preclude the use of a truly clinical analysis. Sarbin, who did better than some of the others, provided his prophets only with high school rank in class, aptitude test scores, a preliminary interviewer's notes, and a paper and pencil personality inventory. Apparently no one, save the "preliminary interviewer," who left only "notes," had looked at the person in action. From such straws the clinician was asked to make bricks! That the clinicians in this study did as well as the actuaries is irrelevant; what they had to be doing, with such non-clinical data, was what Sarbin accuses them of doing: they were managing somehow to function as a human substitute for an I.B.M. machine. Almost all the other studies supply non-clinical data; all demand multiple choice, non-contingent predictions.

A second group of studies includes recent large-scale follow-ups on assessment batteries, such as Murray's O.S.S. program, the Kelly and Fiske studies at Michigan on predicting the success of clinical psychologists, and the California studies of personality that are beginning to be published. None of these has suggested any great validity for the clinical method. We have to take these failures of the clinical method more seriously; they were designed by good clinicians and used excellent clinical data. One presumes that proper clinical analysis got applied, though this is not always clear from the published accounts. What vitiates all these studies, however, is their failure, in two senses, to make clinical predictions. First, there seems to be little or no contingent prediction. Worse, nearly all the predictions take the form of rating scales. That decision in designing these studies determined the nature of the findings.

* * *

So we still don't know the answer to the main question before us.

Only a study under proper conditions will be conclusive. If clinical predictions under ideal conditions fail to come true, running the actuarial half of the experiment will hardly be required! I happen to believe, however, that clinical predictions, as operationally defined in this paper, will turn out to be 100% true; 100%, that is, less only the sampling error that is inevitable because we see 40 hours and not 40 years of our subject's behavior, and less the error arising from unreliability of those who observe both the independent variables and the criterion variable.

That's my null hypothesis. I, like all my fellow clinicians, am eager to see the hypothesis tested.

BIBLIOGRAPHY

1. MEEHL, P. E. *Clinical vs. Statistical Prediction.* Minneapolis: Univ. of Minnesota Press, 1954.
2. SHNEIDMAN, E. S. *Thematic Test Analysis.* N. Y.: Grune & Stratton, 1951.
3. TOMKINS, S. S. *The Thematic Apperception Test.* N. Y.: Grune & Stratton, 1947.
4. WHITE, R. W. "What is tested by psychological tests?" In HOCH, P. H. AND ZUBIN, J. *Relation of Psychological Tests to Psychiatry.* N. Y.: Grune & Stratton, 1951, pp. 3-14.

Clinical vs. Actuarial Prediction: A Pseudo-problem

JOSEPH ZUBIN

There are three possible ways of dealing with the problem presented by the title of this paper: (1) adopt the clinical point of view (2) adopt the actuarial point of view or (3) declare the dilemma to be non-existent. The latter course is the one I have chosen and as a result I expect to get the brickbats from both sides. Clinicians may accuse me of "leaving the field" because of my inability to cope with the dilemma, while actuarians may regard my approach as merely probing the null hypothesis. I feel, however, that the dilemma is in reality a pseudo-dilemma created by the hopefully temporary gap that now separates the clinician from the research worker.

The reason for my position becomes quite clear in retrospect. I began my career in psychology with a statistical net to bag the elusive differences that may exist between abnormals and normals. Disappointment in this undertaking turned me to the study of the individual case. As a result I began to realize that both sides of the coin—the actuarial and the clinical—belong to each other in an inextricable manner. It was not, however, until I began to study the philosophy of science that I could logically resolve the opposition between the two approaches.†

Scientific method is characterized by a continued interaction between observation and schematization (1). Which came first is difficult to determine. Primitive man's observation of nature soon led him to notice certain regularities which he schematized into expectation or hypothesis as we now call it. These hunches, hypotheses or discoveries, if you will, constitute the first step—the context of discovery according to Reichenbach (9). This step might be likened to the storming of a beachhead in the continuing war between science and ignorance. The second step is to verify the hypothesis. This leads us to the context of justification which might be likened to the establishing of law and order in the territory which the beachhead opened up. No amount of beach-storming, however, can conquer a territory, and no amount of empty drill can lead to victory. It is the sequential interaction between the two contexts that leads to success. The clinician, on the one hand, often becomes lost in the land of discovery, narcissistically enjoying every new idea, smelling every new hunch and titillated by every new possibility but only too rarely, if ever, leaving the context of discovery for the context of verification. The actuarian, on the other hand, often becomes lost among his equations, gadgets and techniques, sharpening and polishing under the assumption that the sharper the tool, the better the eventual results. But, for much of our work our tools are already too fine. Most of the concepts which we deal with clinically are too open, too crude to warrant even the .01 level of confidence on the score of either type of inference error (Type I or Type II). But psychology is not alone in this fix. Even biology, a science supposedly higher in the hierarchy of exactness, suffers from loosely defined concepts which nevertheless do not prevent scientific progress. Julian Huxley (4), in defining the concept of species, says:

---

† I owe much of this insight to Dr. Eugene I. Burdock and to Dr. Raymond J. McCall, former students who guided my reluctant steps.

> "However, we must remember that species and other taxonomic categories may be of very different type and significance in different groups; and also that there is no single criterion of species. Morphological difference; failure to interbreed; infertility of offspring: ecological, geographical, or genetical distinctness—all those must be taken into account, but none of them singly is decisive. Failure to interbreed or to produce fertile offspring is the nearest approach to a positive criterion. It is, however, meaningless in apogamous forms, and as a negative criterion it is not applicable, many obviously distinct species, especially of plants, yielding fertile offspring, often with free Mendelian recombination on crossing. A combination of criteria is needed, together with some sort of flair. With the aid of these, it is remarkable how the variety of organic life falls apart into biologically discontinuous groups. In the great majority of cases species can be readily delimited, and appear as natural entities, not merely convenient fictions of the human intellect. Whenever intensive analysis has been applied, it on the whole, confirms the judgments of classical taxonomy."

It is thus not the precision of the concept, but its power in explaining behavior which differentiates the good from the poor concepts (5). The clinician who enchants himself with the brilliance of his discoveries and hunches as well as the actuary who spends his time putting a keener edge on his tools and proudly contemplates their sheen are fanatics who have "redoubled their energies when they lost their goal." For the goal, after all, is the verifiable understanding and prediction of human behavior and to achieve this goal, the observations of the clinicians and his hunches as well as the verification of these hunches by the actuary are essential.

From this point of view, the question of whether the actuarial approach is superior to the clinical is tantamount to asking whether the sperm is more important than the ovum. Both are equally important and no progress can be made with one alone. In fact, exercising one alone in isolation from the other is a rather unproductive form of activity despite the satisfaction it may afford.

The better the hunches, the more effective will be the actuarial prediction, once the hunches are verified. To compare, clinical impressionistic prognosis with the actuarial prognosis derived from a previously formulated clinical hunch is a travesty! How could a new untried clinical impression ever equal the statistically verified residue of earlier clinical impressions. We should have been so certain of our actuarial techniques that nothing but a complete victory in every precinct should have satisfied. Why did the results of the 24 studies (8) fail to show an advantage in each instance. The answer lies in the relative rigor or looseness of the criteria. When rigorous, specific and specified criteria are available, one can always build tests which will prognosticate successfully. As the criteria become looser and less explicit it is debatable whether either method, actuarial or clinical, can accomplish much. Prognoses of mental illness, for example, should be based on a specified follow-up period since outcome varies with period of follow-up. If the actuarial formula is based on immediate outcome as a criterion while the clinical prognosis is based on eventual outcome, it is no wonder that the actuarial method is superior when the results are evaluated against immediate outcome.†

---

† While Professor Meehl did not read his discussion for lack of time, the few remarks he made led me to make the following comments in order to clarify our differing points of view: I had anticipated brickbats from the right and from the left, but not from the center. Despite Paul's very thoughtful book (8), the distinction between actuarial and clinical prediction is heuristic rather than basic. The process of prediction for a group is quite different from prediction for an individual. The former can be completely actuarial as in life expectancy tables; the latter by its very nature must be clinical if it is to result in action. A distinction needs to be made between a *prediction* and a *decision* based on that prediction. The prediction might be that there is a .70 probability of success. What one does on the basis of such a probability in the case of a single individual is best exemplified by what one does for *himself* when faced with such a prediction. In the last analysis, decision is a "clinical" act, not an "actuarial" one. To have one standard in mind when one makes decisions about his own fate, and another standard when one makes decisions for a patient is the "double standard" at its worst. No one would select a secretary or a wife on test scores alone, even if the multiple r were as high as .80 (which it

rarely is in any prediction studies I have seen). Why should one be willing to decide on a patient's therapy on actuarial grounds alone? Mind you, I am not arguing against utilizing regression equations for prediction; but I am concerned with what you do as a consequence of the prediction. When actuarial predictions succeed in encompassing 90% of the variance in the behavior under observation, we can safely leave prediction to a statistical clerk and save the clinician's time for the more arduous task of therapy. Since most actuarial predictions account for less than half of the variance in the observed behavior, actions based on such predictions need the integrating act of the clinical decision.

When the clinician makes a prediction, looks up tables of dosages of drugs, contemplates syndromes of symptoms, he is engaged in statistical or actuarial activity. What he does with this information—his volitional decision—is a "clinical" act.

When the statistician chooses an experimental design, selects a technique or decides on the relative weights to be assigned to certain factors, he is acting clinically. His subsequent analysis and the predictions derived from probability considerations are, of course, actuarial.

The complete process by which a decision is reached with or without the help of Wald's decision functions is a volitional act which has been described introspectively by Ach (Ach, N. "Analyse des Willens". Handbuch der biologischen Arbeitsmethoden, Abt. VI, Teil E., Berlin, Urban and Schwarzenberg, 1935*.

According to Ach, man is never closer to his inner self than when he makes a volitional decision. Freedom of the will, apparent or real, underlies this decision-making process, and is the very essence of mental life. To maintain that in our present state of ignorance, we can substitute a regression equation for the volitional act would be flying in the face of reality. Decision belongs to the context of discovery, a land whose rules and regulations are as yet unknown.

Nevertheless, it is important to call the attention of clinicians to the fact that they have spent too much time in "hunch-land" and not enough in the land of verification. By the same token it is important to indicate to the statistician that the assumptions of normality, linearity, continuity, homoschedasticity, etc., etc. which underlie many of his techniques including the multiple regression equation, discriminant functions as well as factor analysis, are not suitable for the non-linear, discontinuous, unit-less type of observation which the clinician deals with. Between the land of discovery and the land of verification a bridge must be built, consisting of the proper techniques to meet the clinical needs. Clinical psychology today is in about the same position that agriculture was before Fisher or physics was before Newton. Just as Fisher had to develop techniques for dealing with the hunches emanating from the practical agronomist, so a new Fisher is required to develop techniques for testing the hunches emanating from the clinic. This new Fisher will have to convert our present group-centered techniques into individual-centered tools, will have to deal with syndromes and patterns and profiles emanating not from data which satisfy the requirements of factor analysis, but from the crude amorphous qualitative data which defy factor analytic methods, or which are verily disemboweled by such high-powered techniques.

---

* The author translated this book into English some ten years ago and several carbon copies are available on loan.

A good case in point is a recent study on the effects of drugs on psychological test function (7). In order to determine the effect of a new antihistamine on psychological test performance, the effect of the new drug was contrasted with the effect of a placebo, a stimulant and a hypnotic drug. The psychological techniques consisted of a group of conceptual, perceptual and psychomotor tasks and an interview. The results of one of the tests, the critical flicker fusion test, will be sufficient to clarify the point at issue (3). The means of the group of 24 patients who participated in this experiment are shown in Table 1.

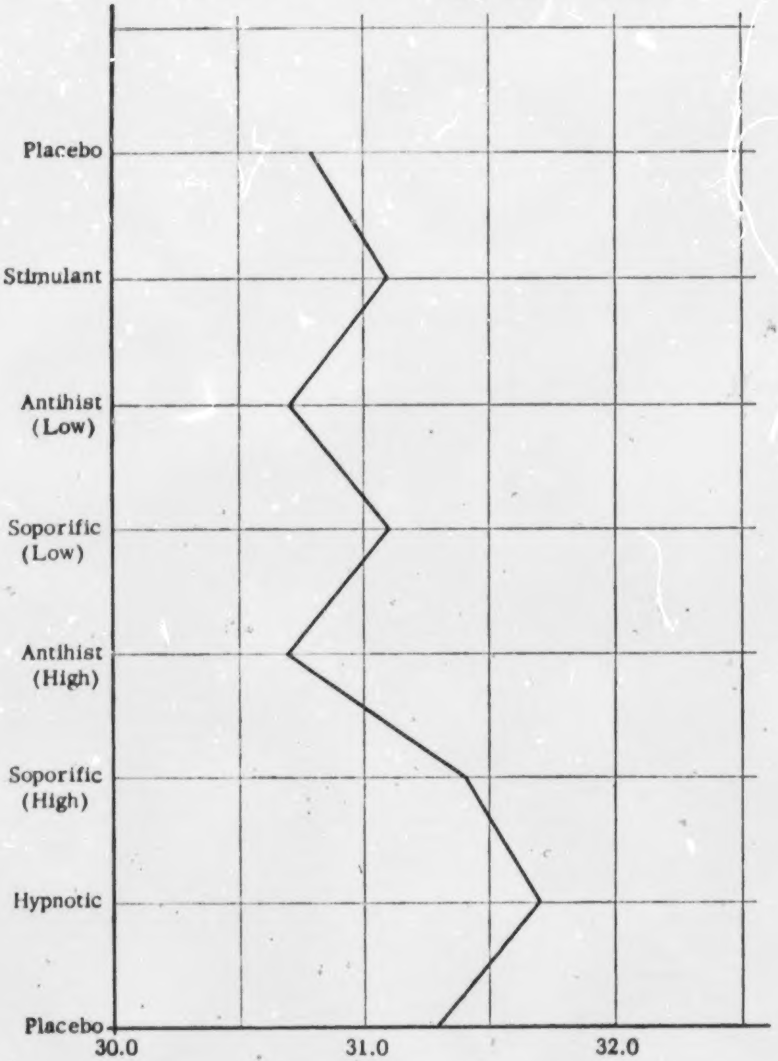
TABLE 1

The critical Flicker Fusion Threshold in cycles per second for the Various Chemical Agents (N = 24).

AGENT	DAY	MEAN
Placebo	1	30.8
Stimulant	2	31.1
Antihist (low)	3	30.7
Soporific (low)	4	31.1
Antihist (high)	5	30.7
Soporific (high)	6	31.4
Hypnotic	7	31.7
Placebo	8	31.3

CHART 1

The Critical Flicker Fusion Threshold in Cycles per Second for the Various Chemical Agents (N = 24)



The data were subjected to an analysis of variance the results of which are shown in Table 2.

TABLE 2

Summary of the results of the total analysis of variance for three threshold determinations of CFF at three levels of apparent brightness at each of the two light-dark ratios for twenty-two subjects over eight days.

Sources of variation	(1) Sums of Squares	(2) %	(3) df	(4) Mean Squares	(5) F	(6) P
1 Between Agents	278.97	0.50	7	39.85	1.83	—
2 " Levels	38572.66	68.67	2	19286.33	118.42	.01
3 " Instruments	2190.67	3.90	1	2190.67	13.45	—
4 " Individuals	7202.85	12.82	21	342.99	15.78	.01*
1 - 2	115.64	0.21	14	8.26	2.12	—
1 - 3	110.24	0.20	7	15.75	2.08	.05
1 - 4	3193.69	5.69	147	21.73	2.87	.01*
2 - 3	325.73	0.58	2	162.87	30.44	.01
2 - 4	370.03	0.66	42	8.81	1.65	.05*
3 - 4	379.38	0.68	21	18.07	2.39	.05*
1 - 2 - 3	54.43	0.10	14	3.89	1.69	—
1 - 2 - 4	979.70	1.74	294	3.33	1.45	.01*
2 - 3 - 4	224.75	0.40	42	5.35	2.33	.01*
1 - 3 - 4	1111.29	1.98	147	7.56	3.29	.01*
1 - 2 - 3 - 4	675.71	1.20	294	2.30		
Within	387.47	0.69	2112	0.18		
Total	56173.21	100.02	3167			

It will be noted that the "between-agent" variance was not significant when compared with the largest interaction term but the "between-individual-variance" and its interactions were statistically significant as shown by the starred F ratios.

Because of the significance of the interindividual variance and its interactions, each individual subject was treated separately as an independent universe. Since 10 measures of critical flicker fusion threshold were taken each day on each individual, an analysis of variance for the single individual could be performed. The results indicated that the group treatment of the data had hidden more than it revealed. The individual treatment of the data indicated that half of the group (11 cases) had remained unaffected by the chemical agents. In those who showed significant effects, the low soporific dosage showed a significantly improved performance in 6 subjects and a significantly poorer performance in two subjects. The higher dosage of the soporific agent improved the performance of 8 subjects and reduced the performance of 3, leaving the other 11 subjects unaffected. The rest of the data are shown in Table 3.

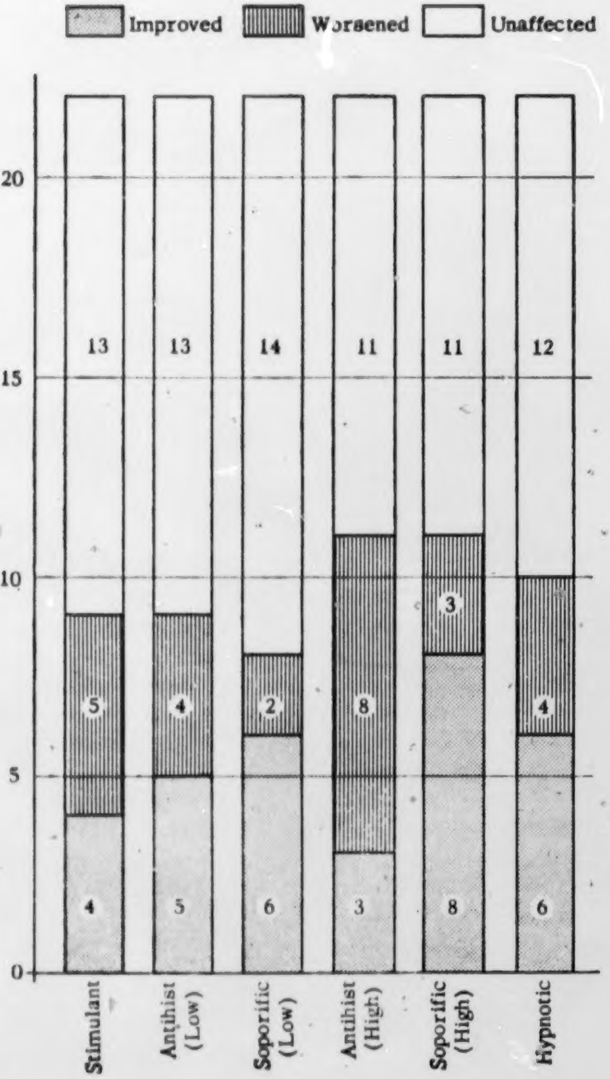
TABLE 3

Number of subjects showing significant improvement or worsening for each chemical agent on the critical Flicker Fusion Test (Stroboscope).

Chemical Agent	Improved	Worse	Unaffected	N
Stimulant	4	5	13	22
Antihistamine (low)	5	4	13	22
Soporofic (low)	6	2	14	22
Antihistamine (high)	3	8	11	22
Soporific (high)	8	3	11	22
Hypnotic	6	4	12	22
Total	32	26	74	132
Average	5.4	4.3	12.3	22

CHART 2

Number of Subjects Showing Significant Improvement or Worsening for each Chemical Agent on the Critical Flicker Fusion Test (Stroboscope)



It is clear that the group of subjects was quite heterogeneous with respect to the effect of the various chemical agents. For this reason group statistics should always be examined in conjunction with individual statistics wherever possible.

Just how a heterogeneous group can be subdivided into more homogeneous subgroups becomes an important question for the clinical-actuarial controversy. If we could find a technique for subdividing a group into homogeneous subgroups, we could then apply group statistics to the subgroups and avoid the impasse which occurred in the previous example.

An example of the application of individual-centered techniques which keeps the sights of the experimenter focused on the individual instead of on the group is the technique of like-mindedness (10). Some 20 years ago we faced the problem of developing a personality inventory which would be of help in classifying mental patients. This study was reported in part in 1937 but because of an error in computation lay uncompleted until recently when the error was discovered and the analysis completed. While we have since given up the use of inventory items as the sole basis for classification, and have (we believe) found more pertinent indicators, the method is general enough to be applicable to most of the data in the clinical field.

The Personality Inventory Form (6) which consisted of a distillate of 70 items from a matrix of 1000 found in other inventories and in case histories, was administered to some 1000 patients of varying types of illness and to 1000 normal controls. In the process of selecting the 70 items, only those items were retained which differentiated the patients from the normals in all the age groupings, the two sex groups, and illness categories, since we wished to get a screening test which would separate the ill from the well. In retrospect this seems to have been a mistake. In picking out only the items which differentiated, we selected the liabilities of the patient group, and eliminated their assets. Perhaps the patterning of the assets and liabilities is a more useful basis for screening than the total number of liabilities alone.

A sample of 68 male schizophrenic patients and 68 normal controls matched for age, sex and education was then obtained and by the use of IBM scoring machines it was possible to obtain the agreement scores of each patient with each of his 67 colleagues and each of the 68 normal controls. Similarly the agreement scores for the normals were also obtained. A sample of the agreement scores is shown in Table 4.

TABLE 4

Agreement Scores between 5 individuals of the abnormal group on a test of 70 items.

	Individuals				
Individuals	A	B	C	D	E
A		37	32	48	44
B	37		47	48	50
C	32	47		46	46
D	48	48	46		53
E	44	50	46	53	

The mean agreement scores are shown in Table 5 and Table 5A and Chart 3.

TABLE 5

Intragroup and Extragroup Agreement Scores for 68 Schizophrenic patients and 68 matched normal controls.

	Intragroup		Extragroup	
Scores	Normal	Schizophrenic	Normal	Schizophrenic
48 - 50	100.0			100.0
45 - 47	70.6		100.0	82.4
42 - 44	45.6	100.0	97.1	73.6
39 - 41	23.5	80.9	55.9	53.0
36 - 38	11.7	60.3	32.4	41.2
33 - 35	7.3	32.4	14.7	26.5
30 - 32	2.9	22.1	8.8	20.6
27 - 29	0.0	10.3	0.0	8.8
24 - 26		5.9		2.9
21 - 23		3.0		2.9
18 - 20		1.5		0.0
15 - 17		0.0		

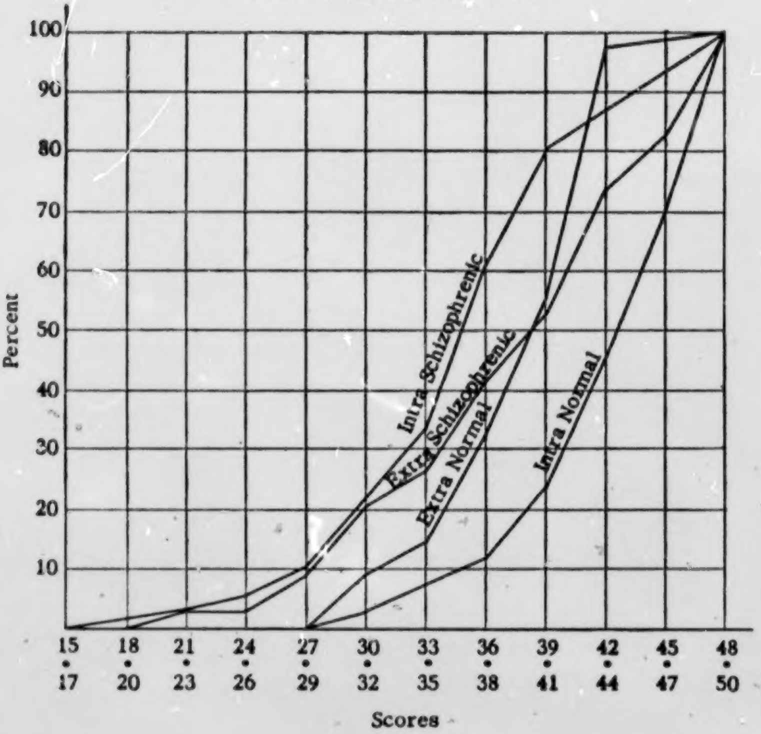
TABLE 5A

Intra-group agreement scores of 68 schizophrenics and 68 matched normal controls and extra-group agreement scores of 34 schizophrenics and 34 matched controls.

	Agreement Scores					
	Intra Group			Extra Group		
Group	N	M	σ	N	M	σ
Normal Controls	68	44.6	4.69	34	40.2	3.94
Schizophrenics	68	37.0	5.44	34	40.2	7.31
Difference		7.6			——	3.37
"t"		8.5				3.34
P		$<.01$				$<.01$

CHART 3

Cumulative per cent distribution of intra-group agreement scores of 68 schizophrenics and 68 matched normal controls and of extra-group agreement scores of 34 schizophrenics and 34 matched normal controls.



The 67 pairs of agreement scores for each pair of individuals were then correlated and the table of intercorrelations of these agreements scores were subjected to a factor analysis. Table 6 shows the intercorrelations.

TABLE 6

Correlation between agreement scores for normals and schizophrenics. (The figures above the long diagonal are for the normals, the figures below are for the schizophrenics.)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
A.		−.011	−.211	−.387	.180	.052	.267	.086	−.008	−.053	.067	−.141	−.512	−.141	−.077	−.220	−.012	1.
B.	.504		.728	.631	.607	.518	.372	.741	.822	.748	.846	.554	.280	.840	.895	.840	.782	2.
C.	.414	.856		.478	.387	.277	.034	.673	.731	.683	.682	.605	.506	.836	.743	.783	.654	3.
D	.701	.863	.839		.322	.190	.055	.389	.467	.571	.482	.409	.553	.537	.588	.580	.518	4.
E.	.647	.864	.839	.920		.571	.570	.570	.759	.421	.653	.383	−.066	.4[illegible]1	.609	.544	.485	5.
F.	.704	.846	.841	.956	.912		.563	.342	.609	.556	.405	.297	−.057	.389	.476	.614	.427	6.
G.	.229	.516	.517	.447	.442	.356		.345	.347	.284	.399	.096	−.241	.198	.407	.286	.376	7.
H.	.396	−.034	.030	.172	.179	.234	.032		.732	.610	.801	.436	.123	.701	.938	.665	.686	8.
I.	.544	.654	.652	.791	.787	.777	.241	.283		.664	.872	.625	.271	.119	.820	.845	.723	9.
J.	.128	−.404	−.366	−.250	−.161	−.174	−.325	.417	.095		.677	−.359	.301	.773	.820	.832	.763	10
K.	.524	.832	.846	.863	.876	.900	.384	.200	.750	−.209		.553	.195	.785	.872	.771	.759	11
L.	.175	.474	.442	.624	.596	.626	.272	.504	.514	−.100	.526		.448	.621	.577	.645	.497	12
M.	.115	−.120	−.170	−.119	.003	−.110	−.091	.164	.012	.343	−.065	−.201		.433	.341	.399	.401	13
N.	.169	−.197	−.171	−.017	.052	−.067	−.271	.276	.345	.568	−.102	.576	.349		905	.852	.670	14
O.	.610	.775	.658	.825	.786	.832	.258	.170	.743	−.086	.773	.627	−.127	.019		.845	.800	15
P.	.325	.014	−.065	.122	.165	.217	−.118	.413	.198	.277	.153	.330	.144	.116	.244		.751	16
Q.	.349	.539	.633	.643	.625	.641	.328	.046	.592	−.199	.645	.181	.221	−.377	.446	.129		17
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	

(I have since been told by Ledyard Tucker that computing the correlations was an unnecessary step, since the agreement scores themselves, after correcting for chance, were a better basis for the subsequent factor analysis.) The factor analyses of the two groups were done separately. Three factors were extracted for normals and four for the patients. These were rotated to simple structure by Dr. R. J. Williams.

The factor loadings are shown in Table 7 and Charts 4A and 4B.

TABLE 7

Loadings on rotated factors underlying agreement scores of 17 schizophrenics and 17 normal controls on the Personal Inventory Form.

ROTATED FACTOR LOADINGS*													
NORMALS						PATIENTS							
Type	Subj.	I	II	III	h²	Type	Subj.	I′	II′	III′	IV′	h′²
I	25	.97	.06	.18	.97	I′	D	.98	−.07	.03	.03	.97
I	19	.92	.00	−.36	.98	I′	F	.97	.07	.05	.15	.97
I	12	.91	.09	.10	.86	I′	E	.96	.06	.05	−.05	.92
I	21	.90	−.07	.19	.85	I′	K	.91	−.07	.10	−.01	.84
I	26	.89	.35	−.01	.92	I′	B	.88	−.21	.05	−.27	.90
I	27	.84	.11	.04	.71	I′	C	.86	−.20	.07	−.30	.89
I	18	.83	−.12	.20	.74	I′	O	.84	.03	.17	.11	.74
I	13	.74	.43	.15	.75	I′	I	.84	.41	.09	−.02	.87
I	15	.73	−.27	.05	.61	I′	A	.68	.17	−.20	.24	.59
I	16	.62	−.13	.14	.42	I′	Q	.67	.08	−.37	.19	.63
I	14	.56	.46	.04	.53							
						{II′	M	−.01	.83	.11	.05	.71
+II	23	.25	.74	.06	.62	{II′	J	−.16	.69	−.03	.29	.59
−II	17	.45	−.52	−.01	.47	IV′	P	.17	.16	.10	.60	.42
+III	22	.48	.32	.68	.79	I′+III′	L	.54	.02	.69	.22	.81
						–	G	.44	.32	.11	−.21	.35
+I+III	24	.75	.29	.58	.98	–	H	.21	−.48	.45	.10	.49
						–	M	−.01	.41	−.38	.09	.32
+I−III	20	.82	.12	−.56	1.00							

* The significant loadings are underscored.

CHART 4 A

Factor Loadings of Type I and Type II Normals on Factors I and II

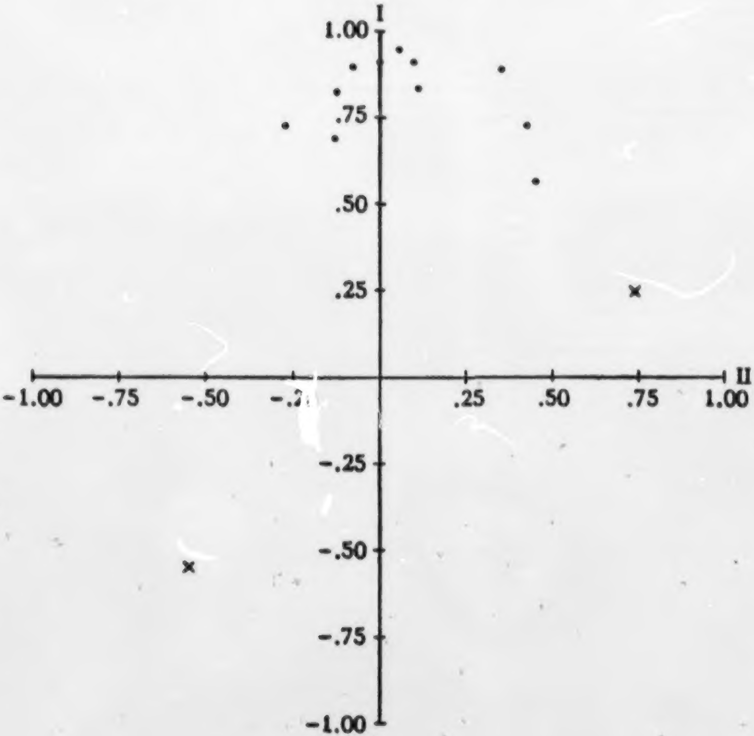
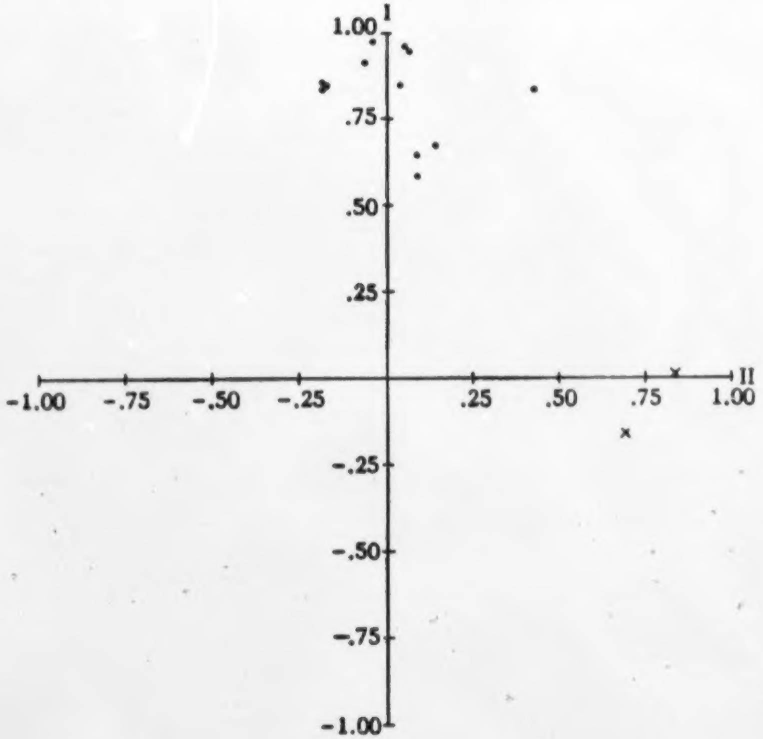


CHART 4B

Factor Loadings of Type I' and Type II' Patients on Factors I and II



Since these factors are merely for the purpose of classifying the patients into homogeneous subgroups, their nature and identity are of no consequence and as soon as we have established the subtypes in our two samples of normals and abnormals—the factors can be discarded. Thirteen of the patients showed a significant loading on only one factor, ten of them on Factor I, 2 on Factor II and one on Factor IV. One patient showed significant loadings on Factors I and III, and three individuals were mavericks, showing no significant loadings on any of the factors. In the normal group, fifteen of the seventeen individuals showed a significant loading on one factor, 11 on Factor I, 3 on Factor II, (one with a positive loading and the other with a negative), one on Factor III, and the remaining two had significant loadings on Factors I and III. It is not profitable to pursue this analysis further except to indicate that this technique permits us to subdivide a large group into like-minded or like-structured subgroups, regardless of the number of variables involved and regardless of the types of distributions that characterize them. It is a type of distribution-free factor analysis. I prefer to regard it as a method for typological analysis. The next step is to find out what the various subtypes have in common and this can be done by studying the common properties of each of the subgroups either with reference to their response pattern or other characteristics such as vital statistics, socio-economic background, genetic factors, etc., etc.

SUMMARY:

I have tried to point out 3 major issues:

1. That the contrast between actuarial and clinical prediction is an unwarranted one. Instead, the two types of prediction supplement each other and the discrepancies between the two should be studied for improving each other reciprocally. Meehl has pointed out that behind the clinician looms the shadow of the actuary and that the latter like the undertaker will have the last word. I doubt this. For behind this actuary is another clinician looking over his shoulder to see just where the formula fails and behind him is a new actuary to see whether the corrections introduced by the clinician hold, etc., etc. I would like to make a plea for the clinician to leave "hunch-space" long enough to see how his hunches hold up and for the actuarian to leave hyperspace long enough to see whether his canonical formulas are applicable and what modifications they need for meeting the demands of the clinic.

2. Secondly, there is a need for more attention to the statistical problem of the evaluation of the individual case. The next break-through in our field is clinical statistics—the gearing of our powerful methods to the consideration of the individual case.

3. Thirdly, there are signs on the horizon that some type of break-through has already taken place. The emergence of interest in pattern analyses or typological analysis is beginning to make a dent in the interaction between clinician and psychometrician. By providing like-minded or like-structured subgroups, it becomes possible to apply present-day statistics to homogeneous groups in our clinical population. This is the first step in the rediscovery of the individual. Our second

most important problem today is to find the pertinent variables for classifying the groups into homogeneous subgroups. Here a reorientation in psychology is called for. But what are the pertinent variables for the description of man? Factor analytic methods have attempted to answer this question. Factor analysis, however, has been applied largely to the conceptual responses of man. The psychomotor, sensory and physiological levels of response have been hardly tapped in factorial studies. But the perceptual and conceptual functions are largely dependent upon man's past experience and to a lesser extent on the immediate "here and now" effects of brain function.

As long as we limit ourselves to the perceptual and conceptual levels, we could regard man as an empty organism. When we begin to examine the behavior of patients we often find that the conceptual area is relatively intact. The functions which have been ingrained in the individual are generally unaltered by shock therapy, psychosurgery and by the disease process itself! The physiological, sensory and psychomotor levels, and the stimulus-bound perceptual level, reflecting as they do immediate brain functioning, are more pertinent for detecting the deviations of the mind. When we develop better techniques for tapping these functions, and apply suitable individual-centered statistical techniques, we may resolve much of the conflict that now exists between the clinic and the laboratory. Just to titillate your appetite for such a classification, the last chart shows a suggested outline (2).

TABLE 8

Examples of measurable activity related to behavior categories and stimulus classes.

	STIMULUS ORDER			
LEVEL OF OBSERVED BEHAVIOR	0 (IDLING STATE) S	I (DISTURBANCES OF HOMEOSTASIS) S	II (INAPPROPRIATE STIMULI) S	III (APPROPRIATE STIMULI) S
CONCEPTUAL	——	ECT; Insulin shock; Lowering of oxygen tension	Electrical stimulation of temporal cortex	Smelling a "sniff set"
R	Reverie and Phantasy	Amnesia, Disorientation, Psychological test performance	Memories Dreams	Recognition of familiar odor
PSYCHOMOTOR	——	ECT	Electrical stimulation of motor cortex	Painful stimulus
R	Spontaneous movement	Seizure	Movement of limb, etc.	Arm withdrawal
PERCEPTUAL	——	Mescal	LSD	Rotating Benham disk
R	Spatial & temporal orientation	Effect on visual orientation	Synaesthesia	Subjective color experience
SENSORY	——	Novocaine	Pressure stimulation above retina; Electrical stimulation of thermal receptors	Light of graded intensity
R	Background noises; cortical gray	Anesthesia	Phosphene Warmth or cold sensation	Threshold response
PHYSIOLOGICAL	——	Hyperventilation	Stimulation by implanted electrodes	Photic driving
R	BMR; Basal EEG; Basal PGR	Effect on EEG	Change in blood steroid pattern	Effect on EEG

TABLE 8 (*Continued*)

Examples of measurable activity related to behavior categories and stimulus classes.

		STIMULUS ORDER		
LEVEL OF OBSERVED BEHAVIOR		IV (CONFIGURAL STIMULI) S	V (SIGNS) S	VI (SYMBOLS) S
CONCEPTUAL		Aircraft forms or silhouettes	Classical delayed response stimuli in animal experimentation	Word association test
	R	Recognition of identity of forms	Successful response by animal subject	Association to stimulus words
PSYCHOMOTOR		Star-shaped maze	Wagging of tail, nuzzling (dog)	Psychiatric interview
	R	Mirror tracing	Petting by human observer	Electromyographic response
PERCEPTUAL		Visual forms	Usual visual alternatives in animal discrimination experiment	Musical tones
	R	Discrimination	Selective response of animal subject	Pitch discrimination
SENSORY		Patterned light stimuli	Infant's faint cry	Words or sentences
	R	Visual threshold	Mother's auditory threshold	Visual threshold
PHYSIOLOGICAL		Patterned visual stimulation	Bell-ringing in Pavlovian conditioning	Verbal instructions to prevaricate
	R	Effect on EEG	Salivation	Effect on PGR

You will note that the left hand column lists the five varieties of responses while the upper row lists the seven types of stimuli which can elicit these responses. Thus, in the idling state, in which no experimental variable is introduced, man is capable of emitting physiological, sensory, perceptual, psychomotor and conceptual responses. Such responses can also be elicited by disturbing man's idling state in some controlled fashion, or by applying an inappropriate or unusual stimulus, an appropriate stimulus, a configural stimulus, a sign stimulus, or a symbol stimulus. Most of our tests have been limited to this upper row—in fact to this last rubric—in which a symbol stimulus elicits a conceptual response. Until we sample this whole table—this behavioral Mendelejeff table—if you will, our understanding of personality, be it of the ill or of the well, will be mighty limited.

REFERENCES

1. Bronowski, J. *The common sense of science.* Cambridge: Mass., Harvard University Press, 1953.
2. Burdock, E. I. and Zubin, J. A rationale for the classification of experimental techniques in abnormal psychology. *J. Gen. Psychol.* (in press).
3. Davis, R. J. A comparison of the stability of the measures of critical flicker-fusion made at two different light-dark ratios as provided by the episocotister and the strobotac. Master's Essay, Columbia University, 1952.
4. Huxley, J. S. Introductory: Towards the new systematics. In Huxley, J. S., Editor, *The New Systematics*, Oxford, 1940.
5. Kaplan, A. Definition and specification of meaning. *J. Philos.* 1946, *43*, 281-288.
6. Landis, C. and Zubin, J. *The Personal Inventory Form.* New York Psychiatric Institute, 1934. (for experimental use only).
7. Landis, C. and Zubin, J. The effect of thonzylamine hydrochloride and phenobarbital sodium on certain psychological functions. *J. Psychol.* 1951, *31*, 181-200.
8. Meehl, P. E., *Clinical vs. Statistical Prediction, A Theoretical Analysis and a Review of the Evidence.* Minneapolis: Univ. of Minnesota Press, 1954.
9. Reichenbach, H. *Experience and Prediction.* Chicago, 1938.
10. Zubin, J. Sociobiological types and methods for their isolation. *Psychiat.* 1938, *1*, 237-247.
11. — A technique for Measuring Like-mindedness. *J. Abn. Soc. Psychol.* 1938, *33*, 508-516.

Clinical Versus Actuarial Prediction

LLOYD G. HUMPHREYS

In the preparation of this paper on clinical versus actuarial prediction* it occurred to me that a third type of prediction might be recognized. I refer to the prediction of responses as a mathematical function of stimulus situation and organism. Whether this constitutes a third case or is to be subsumed under actuarial is of course a matter of definition. Many psychologists would prefer to make this separation. Actuarial prediction would then be restricted, if we use Spence's (4) terminology, to response-response relationships. This is at any rate the class of actuarial prediction with which my paper is concerned. One clinical authority has recently termed this the "engineering" approach. I inferred that he thought of it as a term of opprobrium. I do not find it so and am happy to have this approach referred to as such if you find it meaningful.

In the discussion of clinical prediction I shall restrict myself largely to the situation in which a clinician or counsellor after little acquaintance with the client, with or without test scores, intuitively predicts some future behavior or status for the client. This may not be fair to the clinician but it goes on continually in every clinic and guidance institution. This rules out a second situation, in which clinical predictions are made in therapy while the clinician is gradually forming his hypotheses about the patient. This latter activity is legitimately a human activity and is not to be assigned to a machine. It is also clearly professional in character and is not to be assigned to a clerk. The professional task, however, is to cure the patient; the position of this activity in the development of science is as a source of hypotheses to be tested. It is not a dependable source of knowledge about human behavior.

Before proceeding with the main part of my discussion, it might be pointed out that Meehl (3) underemphasized one important function of the therapist in this second situation. In addition to hypothesis formation on the part of the therapist, evaluation of traits not presently measurable or not well measured—an ability shared with most other people who know the patient well—also takes place. The therapist's ability to predict the behavior of the patient during therapy is in part due to trait evaluation and only in part to the formation of hypotheses about trait combinations and dependencies within the patient. This being true, if I were a clinician—if I may speak hypothetically for the moment—I would want to see my patient in many situations, not merely those involving a couch, in order to obtain maximum breadth of behavior sampling.

---

*I have failed to give individual credit in the discussion to follow to any of my colleagues in the Personnel Research Laboratory (Air Force Personnel and Training Research Center, Lackland Air Force Base) because so many have contributed to both data and ideas and because in a group research organization it is difficult to assign specific credit. My debt should nevertheless be recognized.

With respect to the first situation outlined, I have never had any theoretical a priori expectation that clinical prediction could successfully compete with actuarial predictions. Given valid tests and a valid procedure the clerk and machine should be superior. Fortunately, the evidence surveyed by Meehl supports rather strongly this conviction. The issue is not a serious one, as far as I am concerned, on either theoretical or empirical grounds. This includes Miss Anderson's Employment Service Counseling. In clinical practice, however, it may still constitute an important problem.

It is easy to understand why it is a problem, why it is that attempts are made to second-guess test results by anyone engaged in individual prediction. For this I have no pat dynamic explanation based on the personality structure of clinicians, other than the belief that they are motivated to do a good job. I am referring instead to the size of standard errors of estimate. There is strong motivation here alone to find ways of improving on the information furnished by the best of tests.

What are our hopes of improving on present actuarial predictions by statistical means and thus decreasing the clinician's motivation to do the impossible? There are some obvious things to be done, there are a few things that are perhaps not so obvious, and there are, I am sure, a number of things which are yet to be discovered.

In the first place, we can pay more attention to the reliability of our criteria. There is certainly no point in looking for additional variables or better methods of combination if variability about the predicted criterion score is largely measurement error in the criterion. As a matter of fact we can with clear conscience correct our correlations for unreliability of the criterion in evaluating the quality of the people we place in jobs in a selection or guidance program.

Reliability tells only part of the story. Specificity in the criterion, which is ordinarily considered a part of the reliable variance, is important also. For example, correlations between independent raters concerning a subject's officer quality are substantially higher for a given situation at a given moment in time than if situation and time vary. The degree to which officer quality in general is predictable, however, is a function of the size of correlations between raters when time and situation vary. This is not to say that how a man will be rated by a particular supervisor on a particular job isn't potentially predictable. It does mean that for purposes of evaluating the general trait and for techniques that give us information about the ratee only, this situational specificity should be allowed for. It should also be clear that in order to increase predictions in specific situations we shall need information both about rater and ratee, and knowledge of how this information is to be combined.

We have also been careless about our criterion measures with respect to their homogeneity and comparability for all persons in the sample. Factorial complexity imposes no problem if it is uniform in the sample. But look for a moment at predictions of freshmen grade point ratios in which we typically lump students from all colleges of the university taking dozens of different patterns of courses from further dozens of instructors into a single criterion measure. In addition to the functional complexity of the criterion which varies from one part of the sample to another, we run into problems of lack of comparability of the units of measurement from subsample to subsample. Note that these difficulties with the criterion do not affect its reliability if a student is consistent with respect to his choice of curricula and instructors. I knew one university, for example, in which engineering grades were below the campus average but in which the average engineering student was one standard deviation above the average of the rest of the campus in quantitative ability, two-thirds of a standard deviation above in verbal ability. Over-all correlations with grades were markedly attenuated. A good way to reduce this kind of error is to correlate predictors with separate course grades, obtain intercorrelations of the course grades, and then predict any pattern of courses desired. One typically finds, for example, higher correlations with single course grades than with grade point ratios.

On the test side it is obvious that we need better and additional measures of psychological traits, particularly in the motivation and temperament areas. Our best predictions of later officer quality, for example, are made from personality trait ratings obtained from peers early in training. This does not give us a convenient flexible measure for use in a selection program. One encouraging sign, however, is that we are able to obtain differential validity for trait ratings by these same peers. This finding furnishes both hints and hopes for future test construction.

We have also been looking for additional variables in another, perhaps unusual, way. We have checked comparability of regressions of tests and criteria for different bio-social groups. The typical finding is that, when differences occur, the lines are parallel but the intercepts differ. Females, for example, frequently have higher criterion performance in technical training, test score for test score, than do males. Other differences have been discovered for geographical areas. I do not believe that we should try to adjust for such differences by doing something to the norms. For one thing it is not apparent that these regression differences occur only on tests on which a random sample of females score a compensating amount lower than a random sample of males. Neither is it necessarily true that such regression differences are constant for a given test. It is better to view this finding as evidence that an important variable on which males and females differ has not been measured. Until the variable can be isolated some improvement in prediction can be obtained by weighting sex in the prediction equation.

I won't belabour further the search for additional predictors. A survey of the field would be too time consuming. A less obvious point is that for all kinds of tests we need either a tailor-made job or at least the best possible fit for the group at hand. Not only must the right abilities be measured, but the test must be of appropriate difficulty, with enough items of that difficulty for the group on which it will be used. Appropriateness of difficulty level is not a statistical nicety which makes a difference of .01 or .02 in correlation coefficients. The difference in correlations with outside criteria between using the Armed Forces Qualifying Test, designed to cover the entire range of ability, and a specially designed selection test for a group of officer applicants is measured in the first decimal, not the second. Time limitations on an all-purpose battery are encountered more severely in terms of using sufficient items of appropriate difficulty than in terms of including all the measurable functions necessary.

A major group of problems in prediction can be described in terms of the need for congruence between predictive devices and methods of combination on the one hand and criterion measures on the other. A well discussed example is that the type of process involved, whether additive, conjunctive, or disjunctive, must be comparable for predictors and criteria. Most of the discussion has centered around the applicability of the additive assumption. Actually it seems to be a reasonably accurate model for most kinds of proficiency criteria. Tryout of other models should be most profitable where we have signally failed to date, not where we have been relatively successful. This is not to say that present predictions of academic success, pilot proficiency, or other similar criteria could not be improved through the use of more complex equations than present additive ones. I do believe, however, that gains will be small and difficult to establish. Many pastures are far greener.

Prediction of teaching effectiveness constitutes one example eminently suited for the tryout of other models. Perhaps psychological analysis should have told us this earlier, but the piling up of negative results has clinched the issue. I wonder if perhaps the process involved in this case is disjunctive. Further, I suggest that pattern analysis techniques may be the preferred method of combining variables under this circumstance.

A second example of congruence, or its lack, concerns two additive techniques. Multiple regression is efficient for the prediction of relative success in training or in jobs, but it is not efficient for the prediction of group membership. The multiple discriminant function is an efficient statistic for the latter problem. (It is interesting to note that John French discarded the technique this morning because he selected an inadequate criterion and then brought the multiple discriminate function in again in trying to solve difficulties raised by the use of multiple regression.) Vocational guidance counsellors are generally more concerned with future group membership than they are with potential proficiency. They would make fewer errors in prediction if they could apply the appropriate statistical model. It should be noted that this is not said in a critical spirit—admittedly it will take several years of research and education before we can make effective use of this development.

Two other types of lack of congruence constitute possible sources of attenuation of correlations with criteria. To use Coombs' (1) terminology if we mix relative and irrelative scales, (I would also use ipsative and normative scales interchangeably with Coombs' terms) or if we mix the tasks A and B, set for the subject in being measured, we attenuate the correlations involving such mixed scales.

It appears to me that we have mixed relative and irrelative scales quite indiscriminately. A forced-choice scale of vocational interests is a good example of a relative scale, i.e., measurement is about the subject's own mean. Most proficiency criteria, on the other hand, are irrelative, i.e., measurement is about the mean of the group. If there are large across-the-board differences in interest or motivational level for academic work, we cannot expect to obtain very high correlations between scores on a forced-choice interest test and grade point average. By analogy we certainly wouldn't want to take across-the-board level out of our aptitude battery in predicting this same criterion.

It is of interest to note that types are relative. Somato-type scores add to what is for all practical purposes a constant, i.e., all persons have the same mean. Everyone has a high score some place, no one is low in everything, and there are no persons who are high on everything.

We might describe a perfect type for pole-vaulting, but if a given example of the type were only five foot two he would not be able to vault as high as many faster, taller, and stronger men who did not quite fit the type specifications. Correlations between the type scores and the proficiency criterion would not be as high as a combination of separate measures of height, speed, strength, weight, etc. with that criterion.

It is my impression that clinicians tend to think in terms of types. Perhaps the high and low points in a person's profile are more obvious in the individual interview than his strengths and weaknesses relative to a norm group. It might also be noted that most of the empirical comparisons of clinical and actuarial prediction have involved proficiency criteria. I suspect that some of the astoundingly poor results from clinical prediction result from a combination of relative and irrelative scales.

It seems probable, as we look into this matter further, that there may be some important criteria that are themselves relative. If this were true, relative scales such as those based upon type concepts would predict more accurately than irrelative test scores. I wonder, for example, if perhaps decisions do not involve a balancing of factors within the person more largely than the strength of any one trait or combination of traits, in the normative sense. This problem can still be handled statistically, but we will not find the multiple regression equation which combines results from several irrelative scales very useful.

With respect to the task set for the subject in being measured, it is clear that these should not be mixed and it is possible that they are mixed, willy-nilly, in many situations in which we are trying to predict. Task A of Coombs involves an ideal as the basic frame of reference. The subject is free to select this ideal in many circumstances. Task B involves evaluating a trait or component. For example, if we were to ask a subject to rank 10 politicians in his preferred order, Task A is involved. Presumably he starts with his ideal as rank 1 and the further removed in any direction any politician is from the ideal the lower he is ranked. Now if we ask the subject to rank these same men in their order of liberalism, task B would be involved. Note that the relationship between the two scales resulting from these different tasks is dependent on the position of the subject's ideal on the liberal-conservative continuum. Over many subjects the correlation between the two scales would probably be close to zero. Do we have here a possible explanation for certain low correlations between tests and criteria?

In asking this question, I am not as certain that tasks are frequently mixed as I am that relative and irrelative scales are frequently mixed, but the point is well worth investigating. Even if we ask the subject to assume Task B, it is possible that he will nevertheless be affected by his ideal. Do criterion ratings frequently reflect this phenomenon? Does a consensus of raters merely reflect the average scales obtained from Task A? Needless to say, the low correlations resulting from the hypothetical circumstances would not reflect unfavorably on the tests.

In conclusion, for the situation in which a clinician sees a person briefly and makes intuitive predictions of future status or behavior, I see little hope for the improvement of clinical predictions per se. There is a good deal of improvement possible on the other hand in predictions that we are calling actuarial. This improvement will not take place, however, without a good deal of research. We now have a situation in psychology in which we probably have more tests than there are psychologists doing related research. One of the several important characteristics of this situation is that it allows many degrees of freedom for the operation of chance. I would like to suggest to clinicians that they discard 75% of their test repertoire, perhaps by lot, that they declare a moratorium on the development of additional tests by eager doctoral candidates looking madly for a dissertation topic, and that they concentrate on increasing the complexity of the nomological network, to borrow the terms used by Cronbach and Meehl (2), concerning the tests remaining.

REFERENCES

1. Coombs, Clyde H. *A theory of psychological scaling.* Ann Arbor: Engineering Research Institute, University of Michigan. 1951, 94p.
2. Cronbach, Lee J. and Meehl, Paul E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52,** 281-302.
3. Meehl, Paul E. *Clinical versus statistical prediction: a theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press, 1954, 149p.
4. Spence, Kenneth W. The postulates and methods of behaviorism. *Psychol. Rev.*, 1948, **55,** 67-78.

Clinical Versus Actuarial Prediction

PAUL E. MEEHL

I found Dr. Zubin's empirical data very stimulating; but since they illustrate the use of statistical method in typology and do not bear directly on the predictive efficiency question, I shall not comment upon them further. I am completely baffled by Dr. Zubin's main theme: that the clinical-actuarial issue is a pseudo-problem. I do not find anywhere in his paper a serious attempt at rigorously showing this, and it seems to me that he has clouded the issues by bringing in the interaction between the two methods in *research* work. This research interaction has never been disputed by anyone; all agree that clinicians do generate hunches and, on the other hand, that hunches in social science must usually be *tested* by statistical methods. But the title of this symposium is "Clinical vs. actuarial *prediction*," not "clinical vs. actuarial research-planning." I still maintain that given a finite set of data—tests or otherwise—on an individual patient, for whom a prediction is to be made, you can either hand the data to a clerk or you can hand them to a skilled clinician to think about. Surely this is a pragmatic distinction of real importance. Take a simple, concrete example. We have to decide whether a certain veteran is to be given intensive psychotherapy or not. This is a decision-problem which is being faced in clinics all over the country at this moment.

Does Dr. Zubin seriously assert that we cannot distinguish between these two operations: a naive clerk filling in the values of a regression equation, and 10 clinicians talking around a conference table? Since the latter costs from 10 to 30 times as much (VA rates), Dr. Zubin must have very different notions about economics from mine. *Of course*, the "context of discovery" displays both methods. In my book I emphasized Reichenbach's distinction between the two contexts not once but several times over. In the process of *constructing* a mechanical prediction system, the hunches of clinicians are usually valuable (not always!) and sometimes indispensable. Pick the variables any way you please—using Freudian theory, blind empiricism, or clairvoyance. You may use either "rational" combining functions or choose empirically by blind curve-fitting from a wide class of equations. You may study the hits and misses intensively and qualitatively, hoping to get further hunches as to how the combining function might be improved. At some point, however, you move from the research process to the practical setting; you are asked to apply the fruits of your cerebrations to a realistic prediction problem. *At that moment*, what do you propose to the clinic administrator? Do you give him a statistical table or equation? Or do you tell him to hire a clever psychologist who will think about the same data, *case by case*, and predict therefrom? The first of these solutions is, in daily practice, what I call actuarial, whatever its research history may be. The second solution is non-actuarial, even if actuarial information is part of the total data that the clinician has to "think about." Which of these two procedures has the greater success, the larger hit-frequency, in daily decision-making? This is no academic, hair-splitting question; it is a practical question of intense personal significance to the suffering patient and of great monetary importance to the taxpayer. I find in Dr. Zubin's paper no demonstration that the distinction between clinical and statistical prediction is spurious. Admittedly there are a few borderline methods. But in general, any genuinely *mixed* method is non-actuarial; because the defining property of the pure actuarial method is that it is unmixed. The existence of borderline methods which are difficult to classify does not abolish the distinction (although to believe that it does is one of the commonest of philosophical mistakes). We cannot say precisely how many whiskers it takes to constitute a beard. Any cutting point, as between 78 and 79 whiskers, is arbitrary and subliminal. But we do not conclude that there is no point in distinguishing or that a distinction cannot be made, between a man who is "clean-shaven" and a man who is "fully bearded." Dr. Zubin says the methods "complement each other." This sounds plausible and tolerant; but what does it actually mean? In some of the published studies the effect of allowing the clinician to adjust the actuarial prediction is a shrinkage in predictive efficiency. That seems to me to be a clear case not of complementation but of sabotage. It is senseless to speak of complementation when there are two procedures both purporting to do a specified task but one of these procedures in fact performs the task better than the other, and even better than some mixture of the two procedures will perform it. As to whether a really rock-bottom, epistemological distinction can be made, this is a question of great technical complexity. I would warn everyone against thinking it an easy question, disposable of by a few pleasantries (such as, "the methods complement each other"). Here is needed a thorough analysis using the technical tools of the logicians and mathematicians. I do not know where I stand on this one, and I have spent many hours discussing it with some of the ablest logicians and philosophers-of-science in the business.

Dr. Zubin quotes me as saying that the actuary, like the undertaker, has the final word; and he says he doubts this. He says that the actuary in turn has a clinician looking over *his* shoulder to "see where the formula fails." To which I must reply, so what? At this point, the clinician *thinks* he "sees" where the formula fails; but Dr. Zubin knows as well as I do that this is not the sort of thing you simply "see." We clinicians "see" a lot of things that are not so, if the verb "to see" is used as Dr. Zubin uses it. The context in which I make that remark about the actuary having the "final word" makes sufficiently clear what I mean by this. It is really no more complicated than the scientific principle that I assume we all share, namely, it is facts that check on theories and not the converse. That we will no doubt continue to make still further theories is irrelevant to this primacy of facts; with respect to a *given* theoretical or predictive claim, the facts do have the final word. I can therefore only recommend to Dr. Zubin that he re-read the passage from which he quotes, and ask him to show me specifically where the logic is defective. Jones says that he, using method J, can predict what will happen better than Smith using method S. If Dr. Zubin knows of some way to resolve such a disagreement besides keeping score on Jones and Smith, I should be fascinated to learn what it is. And keeping score—let's be clear about it—is an incurably actuarial process.

Now for Dr. McArthur. I gather he feels there is some kind of disagreement between us, at least with respect to the significance of the available empirical studies. It is perhaps foolish (and not in the symposium tradition!) to say of another scholar's paper: "I agree with everything he says." But I feel impelled to say something very like that about Dr. McArthur. And I don't suppose we can cook up a scientific fight if I insist upon agreeing with him. Let me here say something of a personal nature. I am deeply convinced that in my own therapeutic practice (which is about as psychoanalytically-oriented as one can be without labeling himself a "wild analyst") I do things daily which the best electronic computer cannot begin to do. If I didn't think this, I would feel pretty guilty taking $10 an hour from my clients. I don't see how anyone would even program a computer so as to make it use the raw data as I use them when I interpret a client's dream. It therefore bothers me that clinical psychologists seem to interpret my book as anti-clinical, and pro-statistician; actually, *by far* the larger part of the words in that little volume are devoted to refuting the Sarbin viewpoint. (If you doubt that, just count pages!) At Minnesota we are currently pre-occupied with designing experiments which *are* built to show forth the clinician's unique talents. And I am pretty convinced in advance what the outcome will be; it will be that when a clinician is allowed (quoting Dr. McArthur), to "use the data of his choice, make the analysis of his choice, and make the predictions of his choice," he will look pretty good; not merely better than the actuary, but—more importantly—capable of activities (e.g., open-ended predicting) which the actuary does not even pretend to try. So you see how close I am to the McArthur position. I, like him, believe that we clinikers do special, unique, unduplicable jobs of idiographic conceptualization, when Dr. McArthur's criteria are met by the task and its conditions. *Therefore* I want us clinikers to spend our high-cost time performing these kinds of tasks. Where do we get this time? Well, perhaps there are some other time-consuming activities which we clinikers currently engage in that do *not* meet the McArthur criteria, and in which, consequently, we are at a disadvantage. If the McArthur criteria are applied to perhaps 90% of the prediction tasks which are being daily *attempted* by working clinicians over the country, it is clear that they are not being met. The empirical studies I have surveyed (which now number over two dozen) exhibit a pretty uniform trend. It appears that in prognosis, *given the predictive conditions under which practicing clinicians usually have to operate*, the clinician is largely dispensable or positively adverse to predictive success. Dr. McArthur seems to depreciate the importance of these empirical studies because he sees, quite rightly, that they don't meet his criteria. This puzzles me, because I feel that they are grist for his (and my) clinical mill. (He is wrong about Sarbin, whose clinicians had at least an hour interview with the subjects.) These 25 studies lead me to say, in effect, "Good! Just as I thought, when you don't meet McArthur's criteria, the clinician is beat out by the clerk. So, let the clerk take over these kinds of coarse prognostic and diagnostic tasks. He does it cheaper, and he does it better. I will then occupy my third ear (and Tompkins' souped-up Mill's Methods) with therapy and research." *Part* of this research will be using both methods in a complementary way to develop an equation for the clerk to use. The Harvard Adult Development Study in which Dr. McArthur is engaged I classify as research. If he should propose utilizing the method he describes in the routine predictive tasks of working clinicians, then I will have to start asking him my usual mundane questions about hit-frequency and cost-accounting. Further, Dr. Zubin and I will turn over the McArthur "clinical-introspections" to a super-statistician, just to make sure that with this clinical help in the research context, the actuary is still unable to cook up a mechanical method which will compete with McArthur's clinicians. I, like Dr. McArthur, do not believe that he could; but this is an empirical question. Don't forget—most clinicians would not have expected the uniform trend of the 25 prognostic studies either. But in that non-optimal domain, it seems pretty clear that the clinician's confidence in himself is unjustified by the hard facts. The research task for those who believe, as Dr. McArthur and I do, in the unique clinical powers of the human brain, is to find out whether this belief is true, and in what contexts it is true in a *degree* great enough to be of practical importance.

Drs. Humphreys and Sanford cleverly sent their papers to me after I had already dictated more than fifteen minutes of talk about Drs. Zubin and McArthur. But there is no point anyway in rephrasing their sound and insightful remarks, which is all that I could do. I have a disagreement here and there but it takes too long to develop most of these. I find myself unwilling to agree with Dr. Humphreys' view that we cannot expect to improve those clinical predictions that are based on brief exposure. There is evidence in the literature that people differ in their clinical talents; if we study the process carefully as Dr. McArthur and other researchers (such as Gage, Taft and the IPAR group) are doing, we *should* be able to tease out what is involved in doing it well. In 1944 I checked on the Multiphasic profiles of the patients I chanced to see walking down the hall of the psychiatric unit who appeared to me, at sight only, to be MMPI-psychopaths. During the year I spotted 13 such; in 12 cases I was right. If it were important enough, we could surely learn more about what I was responding to; it must be some fairly crude aspects of dress, appearance, and manner, since I have no psychic powers. And facts about dress, appearance, and manner, once made explicit, are presumably teachable. Dr. Humphreys refers to "pattern analysis" of test scores. Here is a big gap in our knowledge that will not be filled unless you statisticians quit telling us clinicians that Fisher or Hotelling or Rao and Slater solved this problem years ago. They did not. There is to my knowledge no convenient, practical, rigorous procedure for discovering the function and weighting the variables emerging from a many-score test like the Strong, the Multiphasic, or the Rorschach. I will here and now, in the presence of three or four hundred potential takers, offer to name several different clinical problems involving dichotomous criteria in which a Minnesota-trained eye can sort out Multiphasic profiles better than any of these methods. We are currently studying one such Multiphasic task—namely, the discrimination of psychosis from neurosis. I expect the discriminant function to excel the fledgling cliniker, but I expect the skilled cliniker to do still better. Better than all three (and a preliminary study shows this) will be an objective set of complex-pattern rules developed by Dr. Grant Dahlstrom and me. Why am I so confident, a priori, of this order? Because the student clinician follows a near-linear and unconfigured function, non-optimal weights, and low diurnal reliability for identical profiles. The discriminant function eliminates the unreliability and non-optimal weights. The skilled cliniker employs a configural function, and in the case of MMPI this is so important that the superimposed errors of non-optimal weights and unreliability do not wash out the configural gain. Finally, the objective pattern-criteria are configural and the decision is consistent from case to case. Non-optimal weights remain with us. With a 9-variable system, and no underlying theory to suggest a rational combining function, you would have $9+9+36=54$ parameters to fit, if you went past the linear discriminant function to a second-degree expression (with the all-important cross-products). Think, dear brethren, of the sampling errors you would be packing into those 54 constants!

I think Dr. Sanford is right in suggesting that statisticians and clinicians are really interested in predicting different kinds of things. But I want to force this out into the open, because I insist that many working clinicians are blissfully misusing the clinical method to predict the actuary's kind of thing. One program that I am sure all five of us can agree to, and recommend to you as both stimulating and socially significant research, is the empirical study of the two methods of prediction under the various conditions set forth by the four speakers. For what kind of criterion, given what kinds of data, with how much exposure, in what sequence, and so on and on, can the clinician (what clinician?) excel the actuary? There is room for many more studies trying various combinations of conditions before we have the answer. And I should say "answers"; because it will hardly be a decision as to who wins. Rather we will have trustworthy information as to *which* predictive problem is best handled by which method. Here I would like to go into the tremendous matter of *form* versus *content*, which I now tend to see as the real nub of the business. But that would take all night, so it will have to wait for another time.

Appendix

Participants--1955 Invitational Conference on Testing Problems

Adkins, Dorothy C., University of North Carolina
Adlerstein, Arthur, Princeton University
Alexander, Irving E., Princeton University
Allen, Kathryn M., Schenectady Public Schools
Allen, Margaret E., Public Schools, Portland, Maine
Allison, Roger B., Jr., Educational Testing Service
Alman, John E., Boston University
Anastasi, Anne, Fordham University
Anderson, Edward L., Educational Testing Service
Anderson, Gordon V., University of Texas
Anderson, Pauline K., New York State Employment Service
Anderson, Roy N., North Carolina State College
Anderson, T. W., Columbia University
Andree, Robert G., Brookline High School, Massachusetts
Angell, George W., Jr., Educational Testing Service
Angoff, William H., Educational Testing Service
Ansbacher, H. L., University of Vermont
Armstrong, Fred G., Lehigh University
Aronow, Miriam S., New York City Board of Education
Arsenian, Seth, Springfield College, Massachusetts
Bannon, Charles J., Crosby High School, Waterbury, Connecticut
Bardack, Herbert D., New York State Department of Civil Service
Bargmann, Rolf E., University of North Carolina
Barnes, Paul J., World Book Company
Barre, Marguerite F., Vocational Advisory Service, New York City
Barrett, Dorothy M., Hunter College
Bartelme, Phyllis F., New York Regional Respirator and Rehabilitation Center
Bartnik, Robert V., Educational Testing Service
Bauernfeind, Robert H., Science Research Associates
Beck, Hubert Park, City College of New York
Bedard, Joseph A., Public Schools, New Britain, Connecticut
Belt, Sidney L., Educational Testing Service
Bement, Dorothy M., Northampton School for Girls
Benda, Harold W., New Jersey Department of Education
Bennett, George K., The Psychological Corporation
Bennett, Ralph, New York City
Benson, Arthur L., Educational Testing Service
Berdie, Ralph F., University of Minnesota
Bergesen, B. E., Personnel Press, Inc.
Berne, Ellis, New York State Rent Commission
Berrien, F. K., George Washington University
Blackwell, Sara, Educational Testing Service
Blaul, R. Elizabeth, Highland Park High School, Illinois
Bloom, B. S., University of Chicago
Boasi, Veronica M., Department of Personnel, New York City
Boger, Jack Holt, Richmond Public Schools
Boldt, R. F., Educational Testing Service
Bollenbacher, Joan, Cincinnati Public Schools
Bookbinder, Murray, Personnel Department, Philadelphia
Borgatta, Edgar F., Russell Sage Foundation

Bowker, Albert H., Stanford University

Braca, Susan E., Archdiocesan Vocational Service

Brandt, Hyman, American Occupational Therapy Association

Bray, Douglas W., Columbia University

Bretnall, Doris, Educational Records Bureau

Bridgman, Donald S., American Telephone and Telegraph Co.

Bristow, William H., Bureau of Curriculum Research

Brobst, Harry K., Oklahoma A & M College

Broderick, J. Lawrence, YMCA Vocational Service Center

Brolyer, Cecil, New York State Department of Civil Service

Brooks, Richard B., College of William and Mary

Brown, Frederick S., Great Neck Public Schools

Bryan, Miriam M., The Psychological Corporation

Bryan, Ned, Rutgers University

Buckingham, Guy E., Allegheny College

Buel, William D., Temple University

Burdock, E. I., Carnegie Corporation of New York

Burke, James M., Darien Public Schools, Connecticut

Burke, Paul J., Bell Telephone Laboratories

Burnham, Paul S., Yale University

Buros, Oscar K., Rutgers University

Byrne, Lois A., Temple University

Campbell, Donald W., Newark Public Schools

Capps, Marian P., South Carolina State College

Carlson, Harold S., Upsala College

Carlson, J. Spencer, University of Oregon

Carroll, John B., Harvard University

Carstater, Eugene D., Bureau of Naval Personnel

Cayne, Bernard S., Ginn and Company

Chacko, George, Educational Testing Service

Chappell, Bartlett E. S., New York Military Academy

Chauncey, Henry, Educational Testing Service

Christopherson, Helen, Arthur C. Croft Publications

Churchill, Ruth, Antioch College

Cliff, Norman, Educational Testing Service

Cobb, William E., Pennsylvania State University

Cocklin, John H., Temple University

Coffman, William E., Educational Testing Service

Cogan, Blanche, Educational Testing Service

Cohen, Philip S., Montclair State Teachers College

Cole, Joseph W., University of Rochester

Coleman, William, University of Tennessee

Cooper, Hermann, State University of New York

Cox, Henry M., University of Nebraska

Crane, Percy F., University of Maine

Craven, Ethel Case, Polytechnic Institute of Brooklyn

Crawford, Barbara, Educational Testing Service

Crissy, W. J. E., Personnel Development, Inc.

Criswell, Joan H., Office of Naval Research

Cummings, Mary B., Boston Public Schools

Cureton, Edward E., University of Tennessee

Cureton, Louise W., Knoxville, Tennessee

Cutts, Norma E., New Haven State Teachers College

Cynamon, Manuel, Brooklyn College

Dailey, John T., Bureau of Naval Personnel

Daly, Alice T., New York State Department of Education

Damrin, Dora E., Educational Testing Service

Davidoff, M. D., U. S. Civil Service Commission
Davidson, Helen H., City College of New York
Davis, Fred B., Hunter College
Davison, Hugh M., Pennsylvania State University
Day, Robert S., U. S. Military Academy
Dean, E. D. M., Educational Testing Service
Decker, Frederick, Educational Testing Service
Detchen, Lily, Pennsylvania College for Women
Diamond, M. David, Riverside Hospital, New York
Diamond, Lorraine K., Teachers College, Columbia University
Dickson, Gwen Schneidler, Silver Springs, Maryland
Diederich, Paul B., Educational Testing Service
Diers, Helen A., Vocational Advisory Service
Dingilian, David H., Los Angeles City Schools
Dion, Robert, California Test Bureau
Dobbin, John E., Educational Testing Service
Dodds, Alice, Educational Testing Service
Doppelt, Jerome E., The Psychological Corporation
Dragositz, Anna, Educational Testing Service
Drake, L. E., University of Wisconsin
Dressel, Paul L., Michigan State University
Duker, Sam, Brooklyn College
Dunn, Frances E., Brown University
Dunn, Joseph F., Prudential Insurance Company of America
Durost, Walter N., Test Service and Advisement Center
Dubno, Peter, Polytechnic Institute of Brooklyn
Dutton, Eugene, University of Illinois
Dyer, Henry S., Educational Testing Service
Eads, Laura K., New York City Board of Education
Ebel, Robert L., State University of Iowa
Eckert, Ruth E., University of Minnesota
Edelstein, J. David, Educational Testing Service
Edelstein, Ruth R., Bureau of Child Guidance, New York City
Edrington, T. C., Department of Defense
Engelhart, Max D., Chicago Public Schools
Epstein, Bertram, City College of New York
Epstein, Sidney, National Research Council
Estavan, Donald, Educational Testing Service
Evenson, A. B., Department of Education, Alberta, Canada
Fan, C. T., Educational Testing Service
Farabaugh, Mary E., Department of the Navy
Farr, George C., International Business Machines Corporation
Fay, Paul J., New York State Department of Civil Service
Feinberg, M. R., City College of New York
Feldt, Leonard S., State University of Iowa
Fendrick, Paul, Western Electric Company
Fenollosa, George M., Houghton Mifflin Company
Fenstermacher, Guy M., Educational Testing Service
Ferguson, George A., McGill University
Ferris, F. L., Jr., Educational Testing Service
Fifer, Gordon, Test Research Service, Inc.
Findley, Warren G., Educational Testing Service
Fink, August A., Jr., Columbia University
Finkle, Robert B., Metropolitan Life Insurance Company

Fischer, Clyde L., Department of Education, Puerto Rico

Flanagan, John C., American Institute for Research

Fletcher, Frank M., Jr., Ohio State University

Flemming, Edwin G., Burton Bigelow Organization

Fletcher, Carol Ann, Edward W. Hay & Associates, Inc.

Forlano, George, New York City Board of Education

Forrester, Gertrude, West Side High School, Newark

Fox, William H., Indiana University

Freas, Howard J., Jr., Educational Testing Service

Frederiksen, Norman, Educational Testing Service

Freeman, Paul, Educational Testing Service

French, Benjamin, New York State Department of Civil Service

French, John W., Educational Testing Service

Friedenberg, Edgar, Brooklyn College

Friedman, Sidney, Bureau of Naval Personnel

Frutchey, Fred P., U. S. Department of Agriculture

Fulton, Renée J., Bureau of Curriculum Research

Furst, Edward J., University of Michigan

Gallagher, Henrietta L., Educational Testing Service

Gardner, Eric F., Syracuse University

Gaver, Frances, Educational Testing Service

Gelink, Marjorie, The Psychological Corporation

Gerberich, J. Raymond, University of Connecticut

Giangrande, Salvatore C., Cliffside Park Junior High School, New Jersey

Giddings, Frank, Springfield Trade School, Massachusetts

Gise, Helen M., College Entrance Examination Board

Glass, Albert A., The Signal School, Fort Monmouth

Goddard, W. A., International Business Machines Corporation

Godshalk, Fred I., Educational Testing Service

Goldstein, Leo S., Teachers College, Columbia University

Goodman, Samuel M., Puerto Rican Study

Gordon, Mary Alice N., Macy's New York

Graham, Elaine, Bank Street College of Education

Greene, Edward B., Chrysler Corporation

Greene, Paul C., University of Illinois

Gross, Cecily, City College of New York

Grudel, Regina, Teachers College, Columbia University

Guerriero, Michael A., City College of New York

Gulliksen, Harold, Educational Testing Service

Gustad, John W., University of Maryland

Haagen, C. Hess, Wesleyan University

Hagen, Elizabeth, Teachers College, Columbia University

Haggerty, Helen, Personnel Research Branch, Department of the Army

Hagman, Elmer R., Greenwich Public Schools, Connecticut

Hall, Robert G., Manter Hall School, Cambridge, Massachusetts

Halpern, Joseph B., Personnel Department, Stamford, Connecticut

Harmon, Lindsey R., National Research Council

Harper, Bertha P., Personnel Research Branch, Department of the Army

Harter, Roger, American Telephone and Telegraph Company

Hastings, J. Thomas, University of Illinois

Hayes, Rosemary, Educational Testing Service

Healy, Ernest A., Center for Psychological Service, Washington, D. C.

Heath, S. Roy, Jr., Knox College
Heaton, Kenneth L., Richardson, Bellows, Henry and Company
Heil, Louis M., Brooklyn College
Heinemann, Richard F. D., Stewart, Dougall and Associates
Heiser, Ruth Bishop, Goshen, Kentucky
Helmick, John, Educational Testing Service
Helm, Carl, Educational Testing Service
Helmstadter, Gerald C., Educational Testing Service
Hemphill, John K., Educational Testing Service
Herrick, C. James, Rhode Island College of Education
Hieronymus, A. N., State University of Iowa
Hill, Walker H., Michigan State University
Hills, John R., Educational Testing Service
Hirsch, Richard, Educational Testing Service
Hittinger, William F., Haller, Raymond and Brown
Holland, John, Veterans Administration Hospital, Perry Point, Maryland
Hollis, William H., New York City
Hollister, John S., Educational Testing Service
Holley, Clifford S., Personnel Department, Philadelphia
Horton, Clark W., Dartmouth College
Horowitz, Leola S., Queens College
Horowitz, Milton W., Queens College
Howie, Duncan, University of New England, Australia
Hubbard, John P., National Board of Medical Examiners
Huddleston, Edith M., Educational Testing Service
Hughes, J. L., International Business Machines Corporation
Humphreys, Lloyd G., Personnel Research Laboratory, Lackland Air Force Base
Hunt, Thelma, George Washington University
Hunter, Genevieve P., Archdiocesan Vocational Service, New York
Jaspen, Nathan, National League for Nursing, New York
Jeffrey, Thomas E., University of North Carolina
Johnson, A. Pemberton, Newark College of Engineering
Johnson, M. C., Educational Testing Service
Johnson, Theron A., New York State Department of Education
Jordan, Arthur M., University of North Carolina
Kaback, Goldie R., City College of New York
Kalin, Robert, Educational Testing Service
Kalmbach, R. Lynn, Columbia Public Schools, South Carolina
Kaplan, Bernard A., New York State Department of Education
Kelton, John D., University of North Carolina
Kendrick, S. A., College Entrance Examination Board
Keppich, Charlotte, Standard Oil Company (New Jersey)
Keith, A. H., Putnam, Connecticut
Kern, D. W., University of Bridgeport, Connecticut
Kernan, John P., Vick Chemical Company
Kidd, John W., Northwestern State College, Louisiana
Kimball, Elizabeth, Educational Testing Service
Kipnis, David, American Cancer Society
Kleidman, Ruben, Brooklyn College
Kline, William E., The Choate School, Wallingford, Connecticut
Kling, Frederick R., Educational Testing Service
Kogan, Leonard S., Community Service Society, New York
Kogan, Nathan, Harvard University
Kosmerl, Alice, Washington, D. C.
Krathwohl, David R., University of Illinois
Kubis, Joseph F., Fordham University

Kushner, Rose E., City College of New York
Kvaraceus, William C., Boston University
Lambert, Joan, Educational Testing Service
Lamke, T. A., Iowa State Teachers College
Langmuir, C. R., The Psychological Corporation
Lannholm, G. V., Educational Testing Service
Layton, Wilbur L., University of Minnesota
Loughery, Gertrude M., Church Street School, Hamden, Connecticut
Lennon, Roger T., World Book Company
Leverett, Hollis M., American Optical Company
Levine, Richard, Educational Testing Service
Lubtenstein, Ralph, Department of Personnel, New York City
Lindberg, Lucile, Queens College
Lindquist, E. F., State University of Iowa
Litterick, William S., The Harley School, Rochester, New York
Lohman, Maurice A., University of the State of New York
Long, Louis, City College of New York
Lord, Frederic, Educational Testing Service
Lord, Shirley H., Educational Testing Service
Lorge, Irving, Teachers College, Columbia University
Luckey, Bertha M., Cleveland Public Schools
Lusk, Louis T., Norwalk, Connecticut
Lutz, Orpha M. L., State Teachers College, Montclair, New Jersey
Lynaugh, M. B., Western Electric Company
Lyons, William A., New York State Department of Education
Machi, Vincent S., Alfred Politz Research, Inc., New York
MacKay, James L., South San Antonio Public Schools
MacPhail, Andrew H., Brown University
Magoon, Thomas M., University of Maryland
Maltby, Jane M., Board of Education, Hamden, Connecticut
Mandell, Milton M., U. S. Civil Service Commission
Manuel, Herschel T., University of Texas
Marquis, Lloyd K., Arthur C. Croft Publications
Marsh, Mary M., Educational Testing Service
Marston, Helen M., Educational Testing Service
Martin, Harold F., International Business Machines Corporation
Mathewson, Robert H., Division of Teacher Education, New York City
Maxson, Georgia, Educational Testing Service
McArthur, Charles C., Harvard University
McCabe, Frank J., Metropolitan Life Insurance Company
McCall, W. C., University of South Carolina
McCambridge, Barbara, Educational Testing Service
McCann, Forbes E., Personnel Department, Philadelphia
McCully, C. Harold, Veterans Administration
McGillicuddy, Marjorie, New York State Department of Civil Service
McIntire, Paul H., University of New Hampshire
McQuitty, John V., University of Florida
Medley, Donald M., Municipal Colleges of New York
Meehl, Paul E., University of Minnesota
Mellinger, J. J., University of North Carolina
Melville, S. D., Educational Testing Service

Mernyk, Charlotte Levy, Chunky Chocolates
Merwin, Jack C., Syracuse University
Metz, Elliott, New School for Social Research
Michael, Stephen R., Educational Testing Service
Michael, William B., University of Southern California
Micheli, Gene, Metropolitan Life Insurance Company
Mill, Cyril R., Richmond Public Schools
Miller, Howard G., Carnegie Institute of Technology
Millett, Esther, Westover School, Middlebury, Connecticut
Mitchell, Blythe C., World Book Company
Mitzel, Harold E., Division of Teacher Education, New York City
Moll, Clarence, Penn Military College
Mollenkopf, William G., Educational Testing Service
Morgan, Donna D., New York City
Morgan, Henry H., The Psychological Corporation
Morris, Nancy, Educational Testing Service
Morrison, J. Cayce, Puerto Rican Study
Morton, Anton, Educational Testing Service
Mosely, Russell, Wisconsin State Department of Public Instruction
Murray, John E., Special Devices Center, ONR
Myers, Charles T., Educational Testing Service
Myers, Robert L., Temple University
Myers, Sheldon S., Educational Testing Service
Nelson, Kenneth G., New York State Department of Education
Nevin, Margaret, Educational Testing Service
Newman, Sidney H., Department of Health, Education, and Welfare
Nill, Kathryn Fisher, Silver Burdett Company
Noll, Victor H., Michigan State University
North, Robert D., Educational Records Bureau
Nosow, Sigmund, Michigan State University
Olsen, Marjorie, Educational Testing Service
Orleans, Beatrice S., Bureau of Ships, Navy Department
Orleans, Joseph B., George Washington High School, New York City
Orr, David B., Teachers College, Columbia University
Ozkaptan, Halim, Educational Research Corporation
Pace, C. Robert, Syracuse University
Palmer, Harold I., East Orange High School
Palmer, Orville, Educational Testing Service
Patton, James B., Jr., Virginia State Department of Education
Pearson, Richard, Educational Testing Service
Perlman, Mildred, Department of Personnel, New York City
Perloff, Robert, Science Research Associates
Perry, W. D., University of North Carolina
Peterson, Donald A., Life Insurance Agency Management Association
Phillips, Laura M., Silver Burdett Company
Pierson, George A., Queens College
Pinzka, Charles F., Educational Testing Service
Pitcher, Barbara, Educational Testing Service
Plumlee, Lynnette B., Educational Testing Service
Pollack, Norman C., New York State Department of Civil Service
Pratt, Carroll C., Princeton University
Quinn, Edward R., University of Notre Dame
Radasch, John, California Test Bureau
Raine, Walter J., Educational Testing Service
Rapparlie, John H., Owens-Illinois Glass Company

Raskin, Judith G., University of Massachusetts

Reed, Anna K., New York State Department of Civil Service

Regan, James J., Special Devices Center, ONR

Reid, John W., Veterans Administration

Remmers, H. H., Purdue University

Reppert, Harold C., Temple University

Ricciuti, Henry N., University of Colorado Medical School

Ricks, J. H., Jr., The Psychological Corporation

Riessman, Frank, Bard College

Rimalover, Jack K., Educational Testing Service

Rivlin, Harry N., Queens College

Robbins, Irving, Queens College

Ronshaugen, Raydon P., Kent School, Connecticut

Rosenzweig, Allana, Teachers College, Columbia University

Rosinski, Edwin F., University of Buffalo

Rosner, Benjamin, Teachers College, Columbia University

Rulon, P. J., Harvard University

Sait, Edward, Rensselaer Polytechnic Institute

Sands, Elizabeth, Standard Oil Company (New Jersey)

Sanford, Nevitt, Vassar College

Saunders, D. R., Educational Testing Service

Sawin, Enoch I., Air Force ROTC Headquarters, Montgomery

Scates, Alice Y., U. S. Office of Education

Schapiro, Harold B., Knowland & Company

Scheider, Rose W., Educational Testing Service

Schrader, W. B., Educational Testing Service

Schroedel, E. C., International Business Machines Corporation

Schutz, Richard E., World Book Company

Scott, C. Winfield, Vocational Counseling Service, Inc.

Seashore, Harold, The Psychological Corporation

Seibel, Dean W., Educational Testing Service

Sforza, Richard F., New York State Department of Civil Service

Sharp, Catherine, Educational Testing Service

Shaycoft, Marion F., American Institute for Research

Shimberg, Benjamin, Educational Testing Service

Shover, Bertram P., Grosse Pointe University School, Michigan

Silberman, Harry F., City College of New York

Sitgreaves, Rosedith, Teachers College, Columbia University

Slaughter, Robert E., McGraw-Hill Book Company

Smith, Alexander F., New Haven State Teachers College

Smith, Ann Z., Educational Testing Service

Smith, Denzel D., Office of Naval Research

Snodgrass, Robert, Educational Testing Service

Snyder, Betty, Educational Testing Service

Solomon, Herbert, Teachers College, Columbia University

Solomon, Robert, Educational Testing Service

Souther, Mary Tayloe, Tower Hill School, Wilmington

Spaulding, Geraldine, Educational Records Bureau

Spaney, Emma, Queens College

Speer, George S., Illinois Institute of Technology

Spearritt, Donald, Harvard University

Stake, Robert Earl, Princeton University

Stalnaker, John M., National Merit Scholarship Corporation

Stalnaker, Mrs. John M., National Merit Scholarship Corporation

Stecklein, John E., University of Minnesota

Stevens, William C., Veterans Administration Hospital, Perry Point, Maryland

Stewart, Mary, Institute of Physical Medicine and Rehabilitation

Stewart, Naomi, Educational Testing Service

Stice, Glen, Educational Testing Service

Stodola, Quentin, Educational Testing Service

Stokes, Thomas M., Metropolitan Life Insurance Company

Stone, Paul T., Huntingdon College

Stoughton, Robert W., Connecticut State Department of Education

Stovall, F. L., University of Houston

Stuart, William A., Educational Testing Service

Stulbaum, Harold, Metropolitan Life Insurance Company

Super, Donald E., Teachers College, Columbia University

Swanson, Edward O., University of Minnesota

Swineford, Frances, Educational Testing Service

Symonds, Percival M., Teachers College, Columbia University

Tatsuoka, Maurice, Harvard University

Taylor, Justine, Educational Testing Service

Terral, J. E., Educational Testing Service

Thompson, Albert S., Teachers College, Columbia University

Thompson, Kathleen, Educational Research Corporation

Thorndike, Robert L., Teachers College, Columbia University

Thurstone, Thelma G., University of North Carolina

Tiedeman, David V., Harvard University

Tinkle, J. W., Mitchel Air Force Base

Trail, Stanley M., University of Connecticut

Traxler, Arthur E., Educational Records Bureau

Triggs, Frances, Committee on Diagnostic Reading Tests, Inc.

Troyer, Maurice E., International Christian University, Tokyo

Tucker, Ledyard R, Educational Testing Service

Turnbull, William W., Educational Testing Service

Twyford, Loran C., Special Devices Center, ONR

Upshall, Charles C., Eastman Kodak Company

Valley, John, Educational Testing Service

Van Cleve, William J., Educational Testing Service

Vickery, Verna, Southeastern Louisiana College

Viteles, Morris S., University of Pennsylvania

Vose, John C., Educational Testing Service

Wadell, Blandena C., World Book Company

Wagner, E. Paul, Teachers College, Bloomsburg, Pennsylvania

Walker, Helen M., Teachers College, Columbia University

Walsh, B. Thomas, Personnel Department, Philadelphia

Walter, Charles, City of Philadelphia

Walton, Wesley W., Educational Testing Service

Wantman, M. J., University of Rochester

Watkins, Richard W., Educational Testing Service

Watson, Walter S., The Cooper Union

Webster, Harold, Vassar College

Weiss, Eleanor S., Educational Testing Service

Weiss, Joseph, Polytechnic Institute of Brooklyn

Weitz, Henry, Duke University

Wesman, Alexander G., The Psychological Corporation

Whitla, Dean K., Harvard University

Whitney, Alfred G., Life Insurance Agency Management Association

Wilcox, Glenn W., Boston University Junior College

Wilke, Marguerite M., Board of Education, Greenwich, Connecticut

Wilke, Walter H., New York University
Wilks, S. S., Princeton University
Willard, Richard W., Harvard University
Williams, Robert J., Columbia University
Williams, Roger K., Morgan State College, Baltimore
Wilson, John T., National Science Foundation
Wilson, Kenneth M., Princeton University
Wilson, Phyllis C., Queens College
Winans, S. David, New Jersey State Department of Education
Wingo, Alfred L., Virginia State Board of Education
Winterbottom, J. A., Educational Testing Service
Wolf, Beverly, City College of New York
Wolman, Benjamin, City College of New York
Womer, Frank B., Houghton Mifflin Company
Wood, Ray G., Ohio State Department of Education
Wright, Wilbur H., Geneseo State Teachers College
Wrightstone, J. Wayne, Bureau of Educational Research
Zalkind, Sheldon S., City College of New York
Zimiles, Herbert, Bank Street College of Education, New York
Zubin, Joseph, Columbia University